

JPRS Report

Science & Technology

Japan
Mass Data Processing for Genome Analysis

19980506 046

DTIC QUALITY INSPECTED 2

DISTRIBUTION STATEMENT A

**Approved for public release;
Distribution Unlimited**

Science & Technology

Japan

Mass Data Processing for Genome Analysis

JPRS-JST-93-013

CONTENTS

22 March 1993

| | |
|---|----|
| Report of Activities of Research Directors Group [Minoru Kanehisa; MOE KEY AREA RESEARCH, Mar 92] | 1 |
| Planning Research Group: Research on Decentralized Database System for Analysis of Cancer-Related Genes in Human Chromosomes [Akira Ito, Yusuke Nakamura, et al.; MOE KEY AREA RESEARCH, Mar 92] | 2 |
| Knowledge Processing of Sequence Data by Motif Dictionary [Minoru Kanehisa, Atsushi Ogiwara, et al.; MOE KEY AREA RESEARCH, Mar 92] | 4 |
| Construction of an Integrated Database [Satoru Kuhara, Akio Tonouchi, et al.; MOE KEY AREA RESEARCH, Mar 92] | 7 |
| Restructuring of Sequence Data Using Profiles [Osamu Goto; MOE KEY AREA RESEARCH, Mar 92] | 8 |
| Application of an Object-Oriented Model for Management of Hypothesis Knowledge Data [Teruo Koyama; MOE KEY AREA RESEARCH, Mar 92] | 11 |
| Creation of Integrated Database Containing Protein Sequence and Knowledge Data [Yasuhiko Seto, Seiji Itoyama, et al.; MOE KEY AREA RESEARCH, Mar 92] | 12 |
| Organization of Genome Physical Map Data [Eichi Soeda, Xiao-Ren Tang; MOE KEY AREA RESEARCH, Mar 92] | 14 |
| Study on High-Level Searching of Protein Three-Dimensional Structural Data Using Deductive Database Methodology [Toshihisa Takagi, Kenji Sato; MOE KEY AREA RESEARCH, Mar 92] | 14 |
| Research on Determining Amino Acid Sequence Characteristics by Chemical Structure of Ligands [Takaaki Nishioka, Mikita Suyama; MOE KEY AREA RESEARCH, Mar 92] | 17 |
| Evaluation of Parallel Inference Machine in Sequence Analysis and Biological Databases [Katsuki Nitta, Kazumasa Yokota, et al.; MOE KEY AREA RESEARCH, Mar 92] | 19 |
| Research on User Interfaces in Massive Genome Analysis [Masami Hagiya; MOE KEY AREA RESEARCH, Mar 92] | 20 |
| Research on Methods of Expressing Mutation Spectra [Yuzuru Fushimi, Hisashi Sato, et al.; MOE KEY AREA RESEARCH, Mar 92] | 21 |
| Creation of a Database for Experimental Data on Human Genome Analysis and Its Applications [Akisao Fujiyama; MOE KEY AREA RESEARCH, Mar 92] | 22 |
| Structure of Knowledge Base for Transmembrane Proteins [Shigeki Mitaku and Makiko Suwa; MOE KEY AREA RESEARCH, Mar 92] | 26 |
| Structure and Integration of Japanese Language Human Genetic Map (JHGM) Database [Shinsei Minoshima; MOE KEY AREA RESEARCH, Mar 92] | 28 |
| Research on Applications of Super Parallel Computing on Human Genome Analysis [Akinori Yonezawa; MOE KEY AREA RESEARCH, Mar 92] | 30 |
| Public Subscription Research Groups: High Speed General Character String Searches by Intelligent Algorithm and Parallel Processing [Hiroshi Imai; MOE KEY AREA RESEARCH, Mar 92] | 32 |
| Genome Description by Formal Grammar [Yoshiyuki Kotani, Nobuo Takiguchi; MOE KEY AREA RESEARCH, Mar 92] | 33 |
| Research on Selection of RNA Splicing Sites Using Database and Artificial Intelligence [Hiroshi Sakamoto, Kenta Nakai; MOE KEY AREA RESEARCH, Mar 92] | 34 |
| Study of Organizational System for Genome Analysis of Model Organism [Hideo Shinagawa; MOE KEY AREA RESEARCH, Mar 92] | 35 |
| Research on High-Speed Pattern Comparison Algorithm for Base Sequences [Takeshi Shinohara; MOE KEY AREA RESEARCH, Mar 92] | 36 |
| Development of UNIX Shell for Genome Data Analysis [Akira Sueyama; MOE KEY AREA RESEARCH, Mar 92] | 37 |
| Research on High-Level Processing of Protein Amino Acid Sequence Data Based on Pattern Recognition Methods [Yoshimasa Takahashi, Motokazu Kamimura, et al.; MOE KEY AREA RESEARCH, Mar 92] | 38 |
| Application of Case Based Reasoning to Intelligent Processing of Genome Knowledge Database [Takano Terano; MOE KEY AREA RESEARCH, Mar 92] | 39 |
| Description Method for Protein Three-Dimensional Structure by Logical Type Language [Toshiyuki Noguchi; MOE KEY AREA RESEARCH, Mar 92] | 40 |

| | |
|--|----|
| Construction of Major Histocompatibility Complex DNA Database and Development of Visual Software [Hiroshi Hori, Masaaki Matsuura, et al.; MOE KEY AREA RESEARCH, Mar 92] | 41 |
| Research on Acquisition of Knowledge From Mass Quantities of Genome Data by Parallel Learning Algorithm [Satoru Miyano, Setsuo Arikawa; MOE KEY AREA RESEARCH, Mar 92] | 42 |
| Development of Artificial Intelligence System for Genetic Data Analysis Based on Molecular Evolution [Tamio Yasukawa, Ryoichi Kataoka; MOE KEY AREA RESEARCH, Mar 92] | 44 |

Report of Activities of Research Directors Group
92FE0879A Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 1-3

[Article by Minoru Kanehisa, Chemical Institute, Kyoto University]

[Text] The key area research entitled "Research on Massive Knowledge Data Processing Accompanying Genome Analysis" (abbreviated title: Genome Data) is a 5-year plan that began in 1991. The research organization for the first year consisted of the Research Directors Group, a Planning Research Group, and 12 Public Subscription Research Groups.

Research Directors Group (17 persons)

| | |
|-------------------|---|
| Akira Ishihama | Professor, National Institute of Genetics |
| Katsuki Isono | Professor, Science Dept., Kobe University |
| Shunichi Uchida | Manager, New Generation Computer Technology Development Agency (non-profit corporation) |
| Setsuo Osuga | Professor, Research Center for Advanced Science and Technology, University of Tokyo |
| Minoru Kanehisa | Chemical Institute, Kyoto University |
| Yoshiyuki Sakaki | Professor, Dept. of Medicine Laboratory, University of Tokyo |
| Shinichi Sasaki | President, Toyohashi University of Technology |
| Nobuyoshi Shimizu | Professor, Dept. of Medicine, Keio University |
| Yoshiaki Suzuki | Professor, Basic Biology Institute, National Joint Research Agency, Okazaki |
| Mutsuo Sekiguchi | Professor, Dept. of Medicine, Kyushu University |
| Michiru Kora | Professor, Chemical Institute, Kyoto University |
| Hozumi Tanaka | Professor, Engineering Dept., Tokyo Institute of Technology |
| Takaaki Nishioka | Asst. Professor, Chemical Institute, Kyoto University |
| Kenichi Matsubara | Professor, Cell Engineering Center, Osaka University |
| Shigeki Mitaku | Asst. Professor, Engineering Dept., Tokyo University of Agriculture and Engineering |
| Hiroshi Yoshikawa | Professor, Dept. of Medicine, Osaka University |
| Akinori Yonezawa | Professor, Science Dept., University of Tokyo |

Planning Research Group (16 persons)

| | |
|-----------------|---|
| Akira Ito | Manager, Cancer Institute, Japan Foundation for Cancer Research |
| Minoru Kanehisa | Chemical Institute, Kyoto University |
| Satoru Kuhara | Asst. Professor, Agricultural Research Dept., Kyushu University |

| | |
|-------------------|--|
| Osamu Goto | Senior Researcher, Saitama Prefecture Cancer Center |
| Teruo Koyama | Asst. Professor, Science Information Center |
| Yasuhiko Seto | Senior Researcher, Protein Research Promotion Society (non-profit corporation) |
| Eichi Soeda | Asst. Senior Researcher, Gene Bank, Institute of Physical and Chemical Research |
| Toshihisa Takagi | Asst. Professor, Data Processing Educational Center, Kyushu University |
| Takaaki Nishioka | Asst. Professor, Chemical Institute, Kyoto University |
| Katsuki Nitta | Lab Director, New Generation Computer Technology Development Agency (non-profit corporation) |
| Masami Hagiya | Asst. Professor, Mathematical Analysis Laboratories, Kyoto University |
| Yuzuru Fushimi | Professor, Engineering Dept., Saitama University |
| Akisao Fujiyama | Asst. Professor, National Institute of Genetics |
| Shigeki Mitaku | Asst. Professor, Engineering Dept., Tokyo University of Agriculture and Engineering |
| Shinsei Minoshima | Assistant, Dept. of Medicine, Keio University |
| Akinori Yonezawa | Professor, Science Dept., University of Tokyo |

Public Subscription Research Representatives (12 persons)

| | |
|-------------------|--|
| Hiroshi Imai | Asst. Professor, Science Dept., University of Tokyo |
| Yoshiyuki Kotani | Asst. Professor, Engineering Dept., Tokyo University of Agriculture and Technology |
| Hiroshi Sakamoto | Assistant, Science Dept., Kyoto University |
| Hideo Shinagawa | Asst. Professor, Microbial Disease Institute, Osaka University |
| Takeshi Shinohara | Asst. Professor, Data Engineering Dept., Kyushu Institute of Technology |
| Akira Sueyama | Asst. Professor, Engineering Dept., Technological University of Nagaoka |
| Tadao Takahashi | Asst. Professor, Engineering Dept., Toyohashi University of Technology |
| Takano Terano | Instructor, Social Engineering Director, Tsukuba University |
| Toshiyuki Noguchi | Asst. Professor, Science Dept., Nagoya University |
| Hiroshi Hori | Asst. Professor, Institute of Radiological Medicine for Atomic Bomb Exposure, Hiroshima University |
| Satoru Miyano | Asst. Professor, Basic Information Studies Research Laboratory, Science Dept., Kyushu University |
| Tamio Yasukawa | Professor, Engineering Dept., Tokyo University of Agriculture and Technology |

Details of main activities in 1991 are as follows:

(1) Meetings of Directors of Research

The first general meeting of the Directors of Research was held 17 June 1991 to approve the research plans and area activities for the year. We also discussed research exchanges with the field of computer science and the on-site support structure for genome research. The second general meeting was held on 7 January 1992 and reports on research results and area activities were presented. In addition, research plans and area activities for the coming year were discussed. We also discussed a structure for international cooperation with DOE and NIH.

(2) Tutorials

With the goal of deepening mutual understanding between researchers in bioscience and computer science, the Directors of Research hold a series of tutorials as introductory courses in both disciplines each summer. The first tutorial was held 11 and 12 July 1991 at the Kyoto Heian Kaikan Hall, and the following 11 persons gave lectures: Shinogu Kinami (Niigata University), Akira Shimizu (Kyoto University), Hiroshi Noshima (Osaka University), Masahira Hattori (University of Tokyo), Akisao Fujiyama (National Institute of Genetics), Satoru Kuhara (Kyushu University), Ayumi Shinohara (Kyushu University), Hiroshi Imai (University of Tokyo), Katsuki Nitta (ICOT), Toshihisa Takagi (Kyushu University) and Masaki Hirabara (University of Tokyo). There were about 130 participants.

(3) Public Workshops

Another regular activity of the Directors of Research is a workshop open to the public held every winter for the purpose of information exchange between bioscience researchers and computer science researchers. As a continuation of last year's workshop held by General Research B at the stage of preliminary research, the second workshop was held on 9, 10 December 1991 at the Seiryō Kaikan Hall. Guest lectures were presented by Patrick Winston (MIT), David Lipman (NIH), Setsuo Osuga (University of Tokyo) and Kenichi Matsubara (Osaka University). There were also 39 general lectures and other demonstrations. There were about 300 participants.

(4) Newsletter

We published the first edition of the Genome Information Newsletter in August 1991 and the second edition in March 1992. The newsletter was distributed to about 600 people concerned with the MOE Human Genome Program (creative basic research and key area research), the workshop participants, and tutorial participants.

(5) Genome Net

For setting up the wide area computer network "Genome Net" that will be instrumental in creating a genome research infrastructure, NTT dedicated lines were installed establishing links between the Kyoto University Chemical Institute and the Science Dept., University of Tokyo, and between the Kyoto University Chemical Institute and the Osaka Cell Engineering Center on 20 September 1991. On 7 November a line was installed between the Dept. of

Medicine Laboratory, University of Tokyo and the Science Dept., University of Tokyo. Genome Net will be operated in conjunction with the International Science Network, Science Dept., University of Tokyo (TISN).

Planning Research Group: Research on Decentralized Database System for Analysis of Cancer-Related Genes in Human Chromosomes 92FE0879B Tokyo MOE KEY AREA RESEARCH in Japanese Mar 92 pp 4-7

[Article by Akira Ito and Yusuke Nakamura, Physics Dept., Cancer Institute, Japan Foundation for Cancer Research, Komei Sato, Biochemistry Dept., and Masao Kanemori, Health Statistics, Institute of Public Health]

[Text]

Background and Goals

The goal of this research is to develop a data analysis system for efficiently proceeding with an approach that uses LINKAGE analysis of cancer-related genes (oncogenes oncogenic regulatory genes) in human chromosomes and is already underway at the Cancer Institute. To analyze genetic maps based on experimental RFLP (restriction fragment length polymorphisms) data, we must use the CEPH database, which contains reference data on RFLP analysis, and the GDB database, which contains human genetic maps.

We will establish a chromosome data analysis system using LINKAGE software developed at the University of Utah, and we will proceed with R&D on new data analysis. More specifically, we will establish a database and a data search system for experimental results on chromosome analysis of cancer patients. We hope to study new models and analyze experimental data in order to clarify the multi-step process of oncogenesis.

Further, we also hope to proceed with establishment of a computer network system (Genome Network) to support the analysis of data concerning these cancer-related genes.

Details of Study

To realize the above goals, first we will install the existing LINKAGE analysis system and a genetic data (sequence data) analysis system (GCG software), thus establishing a computer system for analyzing experimental data. We will study network hookup methods so we can use genome-related databases in Japan (Dept. of Medicine Laboratory, University of Tokyo; National Institute of Genetics; Institute of Physical and Chemical Research, Tsukuba, etc.) and overseas (CEPH, GDB, etc.), and establish a decentralized network. Further, we will study methods to make the Genome Net easy for biological researchers to use.

Next, at the Cancer Institute we will create a database for experimental data because LINKAGE analysis results and experimental results from chromosome analysis are accumulating rapidly. Experimental data from the group analyzing chromosomes related to breast cancer (chromosome loss LOH [loss of heterozygosity], oncogene discovery) have been input as tables on a Macintosh, so we will

transfer this to a workstation (VAX + Sun) and make a relational database (SYBASE).

We will attempt a multi-step analysis of the occurrence and progression of cancer by studying the extent of breast cancer malignancy (tumor size, histopathological classification, metastasis to lymph nodes) through the relationship between the loss of regulatory gene groups on specific chromosomes (Nos. 3, 16 and 17) and amplification of an oncogene (erbB-2).

Results

1. Establishment of Chromosome Analysis and Genetic Data (Sequence Data) Analysis Systems

We established LINKAGE analysis and a genetic data analysis system, and we set up a structure for their use on the laboratory local access network. Figure 1 shows a

schematic drawing of the decentralized network system for chromosome and genetic analysis both inside and outside the Cancer Institute. Inside, LINKAGE analysis on the VAX and sequence data analysis by GCG software can be performed freely in each lab on PC terminals that are connected to the internal local access network. Outside, we can access the desired remote computers and databases shown in the drawing by using public packet communications networks. We expect the Genome Network to be hooked up during 1992. Further, we are proceeding with setup of our newly installed SUN workstation and SYBASE database.

This system operates the whole year and is used quite often. It greatly contributes to progress in the creation of chromosome genetic maps and sequence data analysis.

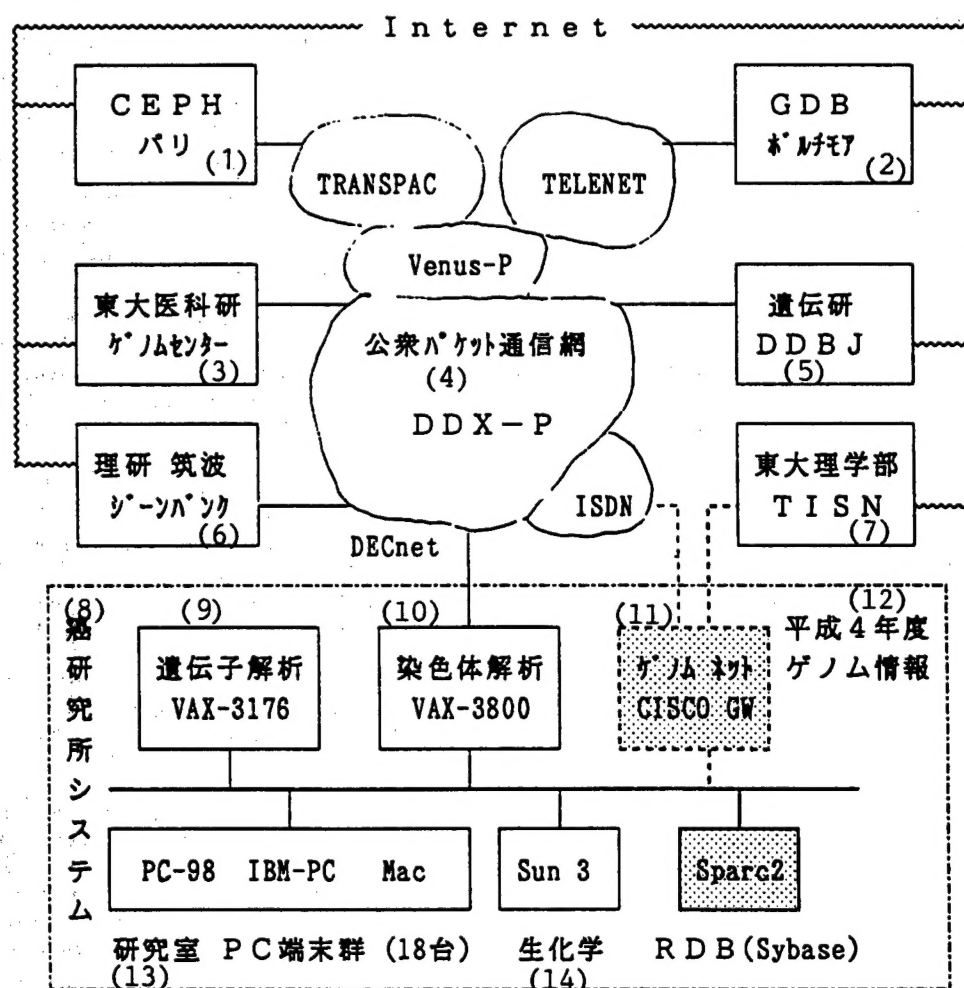


Figure 1. Computer System for Cancer-Related Gene Analysis of Human Chromosomes

Key: 1. CEPH Paris; 2. GDB Baltimore; 3. Genome Center, Dept. of Medicine Laboratory, U. of Tokyo; 4. Public packet communications network DDX-P; 5. National Institute of Genetics DDBJ; 6. Institute of Physical and Chemical Research, Tsukuba; 7. Science Dept., U. of Tokyo; 8. Cancer Institute System; 9. Gene analysis VAX-3176; 10. Chromosome analysis VAX-3800; 11. Genome Net CISCO GW; 12. 1992 Genome data; 13. Lab PC terminal group (18 units); 14. Biochemistry

2. Advancement in Breast Cancer-Related Chromosome Analysis

At the Cancer Institute we are proceeding with genetic analysis of chromosomes associated with the cancers of several organs, but we are focusing on breast cancer, where we continue to make progress in identifying the genes, and proceeding with data analysis. We have analyzed the chromosomes of cancer cells and normal cells from breast cancer patients (219 persons) who have had surgery at the Cancer Institute Hospital. When we investigated chromosome loss with LINKAGE analysis, we found LOH (loss of heterozygosity) in the short arm of chromosome 3 (3p), the long arm of chromosome 13 (13q), the long arm of chromosome 16 (16q) and the short arm of chromosome 17 (17p) in 39 RFLP markers. Further, there was a strong correlation between LOH and the clinical and histopathological classification of tumors (tumor size, metastasis to lymph glands). It became clear that the accumulation of genetic changes in chromosomes contributes to the progression of breast cancer. There was also a strong correlation between LOH on chromosome 17 and amplification of the *erbB-2* oncogene. A detailed genetic map was created for chromosome 3 (Refs. 2, 5). Based on the LINKAGE analysis of DNA data of 40 pairs of CEPH reference pedigrees (total 504 persons), 41 RFLP markers were mapped on Chromosome 3. Creation of this kind of detailed genetic map (Ref. 4) is vital research for not only the clarification of cancer-related genes but also for the Genome Analysis Project.

Discussion

1. Chromosome and Genetic Analysis System Utilizing Existing Analysis Software

The Cancer Institute's chromosome and genetic data analysis system shown in Figure 1 are often used for LINKAGE analysis and DNA/amino acid sequence data analysis. During the 8 months between April and December 1991, LINKAGE data analysis was performed about 500 times, using 260 hours of CPU time. Including sequence data analysis (homology search on FASTA takes about 30 minutes), this was about 20% of the annual CPU (VAX-3800, 3.8 MIPS) usage rate. If the second VAX (VAX-3100/M76, 6.6 MIPS) is also used when needed, we believe our present needs for chromosome and genetic data analysis using the existing software can be met. Genome analysis is progressing, however, and we believe that we will require a new means to do physical mapping in the future.

2. Use of Wide Area Networks and Genome Net

We presently access outside computers and databases via public packet communications networks (DDX-P/Venus-P), and we accomplish this via a link to DECnet (Dept. of Medicine Laboratory, U. of Tokyo; Institute of Physical and Chemical Research, Tsukuba; CEPH on each VAX) and an X.29 terminal link (National Institute of Genetics; GDB, etc.) We carry out relatively smooth data exchange among the genetic data analysis systems, mainly on our present VAX. In the future, we wish to participate

in Genome Net based on the more open Internet for data exchange. Further, we wish to plan for research support so that small research facilities can also participate in Genome Net easily by using an integrated services digital network.

3. Database for Chromosome Analysis Data

We have created a database for breast cancer-related chromosome and genetic analysis data, and for patient clinical and pathological data, and we have made progress in constructing a system that will enable data analysis from many different angles. We are transferring breast cancer data that was input on a Macintosh to a workstation and are creating a relational database. By doing so, we hope to be able to search the various kinds of data freely.

4. Analysis of Multi-Step Process of Oncogenesis in Breast Cancer

The breast cancer research group at the Cancer Institute is diligently pursuing the isolation and identification of the oncogenic regulatory genes (families) for breast cancer and on the clarification of the oncogenic process. We hope to continue our analysis while thinking about models (for example, a neural network) for supporting this approach that employ knowledge data processing.

References

1. Ito, Akira, DNA Detabanku, Rinsho Byori, Tokushu 85 go, 24-31, (1990) [DNA Data Bank, Clinical Pathology, Special Issue Vol. 85].
2. Sato, T., Tanigawa, A., Yamakawa, K., Akiyama, F., Kasumi, F., Sakamoto, G., and Nakamura, Y., Allelotype of Breast Cancer: Cumulative Allele Losses Promote Tumor Progression in Primary Breast Cancer, *Cancer Research*, 50, 7184-7189, (1990).
3. Sato, T., Saito, H., Morita, R., Koi, S., Lee, J. H., and Nakamura, Y., Allelotype of Human Ovarian Cancer, *Cancer Research* 51, 5118-5122, (1991).
4. Yamakawa, K., Morita, R., Takahashi, E., Hori, T., Lathrop, M., and Nakamura, Y., A Genetic Map of 41 Restriction Fragment Length Polymorphism Markers for Human Chromosome 3, *Genomics* 11, 565-572, (1991).
5. Sato, T., Akiyama, F., Sakamoto, G., Kasumi, F., and Nakamura, Y., Accumulation of Genetic Alterations and Progression of Primary Breast Cancer, *Cancer Research* 51, 5794-5799, (1991).

Knowledge Processing of Sequence Data by Motif Dictionary

92FE0879C Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 8-11

[Article by Minoru Kanehisa, Atsushi Ogiwara, and Ikuro Uchiyama, Chemical Institute, Kyoto University]

[Text]

Background and Goals

A homology search, which is used for interpreting the biological significance of sequence data, compares each

item of data in a massive database. Because we have learned through experience that similarities in sequence data reflect similarities in biological function, if a protein with a similar sequence can be found within the database, we can gain some insight into the function of the protein in question. What we must take note of here is that the computer is only used for searching out knowledge from the database, and that making inferences about function based on search results depends on the knowledge and skills of experts. That is because a sequence database is merely a collection of data, and search results are in a form that only humans can interpret. To computerize this process, including the inference-making process, we must construct a knowledge base to serve as a foundation. We are creating a motif dictionary as a knowledge base for interpreting the biological significance of sequence data. A motif is a sequence pattern that is characterized by a specific functional group, and by its presence or absence we can determine whether a protein belongs to a specific group or not (which corresponds to a homology search for group units rather than individual sequence units) and decide what kind of significance that group has (which corresponds to a non-computerized inference process using a homology search). In 1991, we established a process to automatically create a motif dictionary from a sequence database.

Method

(1) Sequence Database and Superfamily Classification

We used Release 26.0 of the PIR protein amino acid sequence database. For sequence data groups, we used the PIR superfamily classification. A superfamily contains groups with similar sequences that are thought to be linked through molecular evolution. The classifications were set up beforehand for the matches to be recognized in the homology search. The present PIR database is divided into three sections, but only the first part (PIR1) is classified into superfamilies, so we only applied our motif extraction procedure to this part.

(2) Unique Peptide Dictionary

The motif extraction procedure is divided into three steps. In the first step screening is performed for peptide patterns (unique peptides) that are characterized by specific groups based on the two conditions of uniqueness and conservation. When a pattern appears only in a specific group, we call that pattern unique. When a pattern has all the sequences within the group, we say that the pattern is conserved. In actual practice, the computer checks the appearance frequency of all oligopeptide patterns 4, 5, and 6 residues long within each superfamily in the database, and it is set up so that conservation is 70-80% rather than 100% so that it can detect patterns with some substitution as well (parameter f).

(3) Consensus in Lining Up Unique Peptides

In the second step, for each group in which a unique peptide was found, we sought the location of each pattern by checking every sequence within the group. We used the sequence data only as data for the distance between one pattern and another to determine the way these patterns line up in the group as a whole. Then, just as if we were aligning multiple sequences, we combined this with a pairwise dynamic programming method to find the alignments.

(4) Making the Consensus Precise

The consensus obtained above is expressed as the maximum and minimum spacers between blocks of uniquely conserved patterns (called motif blocks).

<Motif block1> [Min_spacer, Max_spacer] <Motif block2> ...

In the third step, these are compared once more with the individual sequence data in each group to see whether each block is present or not, and if present, whether amino acid substitution has occurred or not. We set it up so that roughly 70-80% of the amino acids must match for a block to be considered present (parameter r).

Results

(1) Construction of Motif Dictionary

In release 26.0 of the PIR1 database, 7,235 sequences with a total of 2,221,416 residues are classified into 2,350 superfamilies. A certain number of members are needed in groups to extract a motif, and there were only 521 and 283 superfamilies with more than 3 or 5 members respectively. Although it depended on the extent of conservation f, we were able to define motifs in more than half of these relatively large superfamilies. For example, when f = 80% for superfamilies with 5 or more members, we found motifs in 145 of the 283 superfamilies. Thirty-five of the 145 had a single motif block, and the remainder were constructed of multiple motif blocks. In addition, we constructed the motif dictionary with a substituted pattern recognition of r = 80%. A part of this dictionary is shown on next page.

Note: Superfamily number (in brackets), superfamily name, motif pattern (shown on next page)

(2) Use of Motif Dictionary

Next we tested for superfamily affiliation of given sequence data using the motif dictionary with the following procedure. Beginning with the first block, we determined whether a block that r or more amino acids matched was present within a given area or not. If even one block was present, we said that the protein belonged to that superfamily. When we made a prediction using the sequence data of Release 26.0 of the PIR1 database that we used in creating the dictionary, we correctly determined

< 14 > cytochrome b5
-HPGGEEVL

< 31 > L-lactate dehydrogenase
-PVD{I|V}L=[47,47]==G{E|Q}HGD

< 50 > glyceraldehyde-3-phosphate dehydrogenase
+GFGR{I|-}GR=[129,134]==SNASCTTN{C|S}LAP=[14,14]=={L|M}MTTVH=[30,31]=
+TGAA{K|R}A{V|T}=[92,95]=={S|A}WYDNE

< 60 > acyl-CoA oxidase
-TVGDIG=[21,21]==RFFM=[153,159]==ACGGHG

< 68 > glutamate dehydrogenase (NAD(P)+)
+AEG{A|S}N=[24,31]==N{A|C}GGV

< 83 > NADH dehydrogenase (ubiquinone) chain 2
+LS{L|M}GGLPP

< 96 > cytochrome-c oxidase polypeptide I
-{Q|E}HLFWFFGHPEVYI=[126,127]==-VV{A|G}HFHYVLS

< 97 > cytochrome-c oxidase polypeptide II
+G{H|F|Y}QWYW=[83,91]==-YGQCSE{I|L}

< 98 > cytochrome-c oxidase polypeptide III
-SPWPL=[111,113]==PLLNT=[105,106]==-YWHFVDV

< 123 > superoxide dismutase (Cu-Zn)
-HFNP

1,480 items, but 79 that should have been selected were omitted, and 70 items that were not correct were selected.

(3) Motif Block Function, Structure, Evolutionary Significance

Because the extracted motif blocks are conserved sequence patterns within functional groups, we can assume that they correspond to functionally important positions. Further, we can assume that multiple motif blocks that are separated on a sequence exist in near proximity three dimensionally and form functional sites. This was actually confirmed from a detailed analysis of individual superfamilies. In addition, when we studied the correspondence between motif blocks and the exon/intron structure of the DNA base sequence, in 15% of the cases where we know the intron location, there were corresponding motif blocks. In other words, a rather large number of introns lie between functionally important conserved sites.

Discussion

The greatest problem in the creation of a motif dictionary is the fact that the number of superfamilies from which we could extract motifs is a small 145. To increase this number, we attempted to include those superfamilies with a small number of members and lower the conservation ratio f . When $f = 70\%$ and the superfamily has 3 or more

members, we could extract motifs from 324 superfamilies. We also attempted to relax the conditions on uniqueness to allow a small number of exceptions, and to extract unique items from multiple superfamilies rather than unique items from a single superfamily. There are many problems with superfamily classification. Especially, when a protein has multiple domain structures, one sequence must not be classified into just one superfamily. We must also reconsider grouping, which is the preliminary step to motif extraction.

The Human Genome Analysis Center at the Dept. of Medicine Laboratory, University of Tokyo and the Super Computer Laboratory at the Chemical Institute, Kyoto University are planning many services through the wide-ranging computer network "Genome Net." We plan to make the motif dictionary we have created available through Genome Net as well. Further, FASTA and BLAST, which perform homology searching on electronic mail, are also available on a trial basis.

References

1. Ogiwara, A., Uchiyama, I., Seto, Y., and Kanehisa, M., Construction of a dictionary of sequence motifs that characterize groups of related proteins. Submitted for publication.

2. Ogiwara, A., Uchiyama, I., Kanehisa, M., Ruiji Tampakushitsu Gurupu O Tokuchozukeru Hairetsu Mochifu Jisho No Jidosakusei, Dai 2 Kai Genomu Joho Wakushoppu [Automatic creation of a dictionary of sequence motifs that characterize groups of related proteins, 2nd Genome Data Workshop], Tokyo (1991).

3. Ogiwara, A. and Kanehisa, M.; Koiki Nettowaku Genomu Netto No Kozo To Bunsanshori Kankyo No Moto De No Iden Johoshori, Dai 2 kai Genomu Joho wakushoppu, [Genetic data processing based on the structure of the wide-ranging genome network Genome Net and a decentralized processing environment, 2nd Genome Data Workshop], Tokyo (1991).

Construction of an Integrated Database

92FE0879D Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 12-15

[Article by Satoru Kuhara, Akio Tonouchi, and Kyoko Takiguchi, Agricultural Research Dept., Kyushu University, and Emiko Furuichi, Fukuoka Women's Junior College]

[Text]

Background and Goals

It is essential in genome analysis to extract biological data by knowledge data processing and to construct databases. The problems in these fields include the creation of genetic maps and their databases, creating databases for gene base sequences, extracting information from these base sequences (particularly in forming assumptions about the coding regions), creating a database of protein amino acid sequences, predicting protein higher-level structures from amino acid sequences, creating a database of protein tertiary structures and predicting protein function. It will be impossible to solve these problems with existing computer systems, and we must use knowledge data processing methods that have been developed recently. The construction of computer systems incorporating this new methodology will not only speed progress in genome analysis, but it will also become an origin for new research in the field of information science, particularly in the applied field of knowledge data processing technology. Among the above problems, those that especially must incorporate this new method of data processing technology include the extraction of information from base sequences, predictions of higher protein structure and predictions of protein function. Therefore, this integrated database system for genome analysis must include these new methods and become a system that integrates databases of chromosome maps, base sequences, amino acid sequences, and protein tertiary structures.

Details of Study

This year, as the first step in constructing an integrated database, we constructed a hypermedia-type Hyper Genome System that integrates human chromosome maps and base sequences of genes located on the map. Basically, the Hyper Genome System employs a graphic user interface in which details of an item are displayed when the item is extracted by using a pointer. For data sources we used the Genome

Database (GDB) for the human chromosome maps and GenBank for the base sequences. The data in the Hyper Genome System includes LOCUS, SOURCE, and PROBE from GDB and ENTRY from GenBank in its entirety, and we extended a link to each item that related to LOCUS. More specifically, for GDB LOCUS and GenBank ENTRY, we collated the ACCESSION NUMBER listed in GDB PROBE with MAP and GENE listed in GenBank FEATURES and linked them. We used this data as a base, and provided it with system functions such as screen startup, LOCUS screen startup, PROBE screen startup, REFERENCE screen startup, SEQUENCE screen startup, KEYWORD screen startup, CHROMOSOME screen startup, SUPERIMPOSE screen startup and the like.

During this integration, we also studied a system for extracting data from the database. More specifically, with the goal of clarifying the correlation between protein structure and function, we revised the protein high level structure database PACADE. This year, in addition to last year's tertiary structure, we included super-secondary structure, which allows searching for spatial conformations in secondary structure. From the Protein Data Bank, which was the source of the data, we extracted secondary structures, and for helix and sheet structures we determined parallel or antiparallel relationships, calculated the shortest distance between each structure, and added this new information to the database. On this system we created rules for several types of structures and searched the data.

Finally, we developed a methodology for extracting functional sites from primary structure. For this problem we incorporated a machine learning process. In other words, the process creates two sets of data, one sequence with a functional site and another without a functional site, and it automatically determines the characteristics for distinguishing between these two sets. We created a new method in which a decision tree can use a pattern for a node. With this process, it is possible to extract characteristics not only from locations that have functions, but also from locations that clearly do not have functions as well.

Research Results

In the Hyper Genome system we recorded the GENE SYMBOL from 4,000 entries in GDB and 1,500 sequences from GenBank. In this manner we listed the majority of GENE SYMBOL recorded in the chromosome maps for the human genome. As shown in Figure 1 [not reproduced], a person using the system can conduct searches rather easily through the use of a pointer from its graphic user interface.

Furthermore, at the same time we are trying to incorporate data on the mouse chromosome from a mouse genome database. This system is constructed on a SUN workstation.

Next, we carried out an analysis of hydrophobic clusters in proteases on the PACADE database system used for protein high level structural analysis. As a result, we confirmed that although the amino acid sequence homology in three acid proteases (3APP, 4APE, 2APR) is only about 36.9%-53.4%, there is a striking hydrophobic cluster homology as shown in the figure [not reproduced].

We believe that this information will play a major role not only in the clarification of acid protease function, but also in protein modeling, which is part of protein engineering.

Finally, with respect to the development of a methodology for predicting functional sites, it became clear to us that the use of decision trees is very effective. Further, we learned that by simultaneously evaluating for both sets of sequences containing functional sites (positive sets) and sets of sequences that do not have functions (negative sets), we can extract the characteristics of functional sites more accurately. More specifically, we believe that the extraction of negative motifs in research on transmembrane proteins by machine learning has made it applicable to this field of study as well.

Discussion

The construction of an integrated database for genome analysis has just begun, and we are presently at the stage of constructing and testing prototypes. It will be necessary to create specifications for an entire hypermedia type system. Particularly, we must consider the creation of a navigator for genome analysis, which makes great use of graphics for researchers in the biological sciences. The schema of the database must be designed so that data about each mapped gene or base sequence probe, the literature and other data concerning them, data about protein amino acid base sequences that have been coded, and data on high level structure and biological function can all be extracted from chromosome and genetic map graphics. In line with these specifications, this year we constructed the Hyper Genome System, which is one such prototype. We must make trial runs on this system and improve it in the future.

As a result of searches using a protein tertiary structure database system with a deductive inference engine, we learned that it is possible to search for structural characteristics of protein tertiary structure by writing simple rules. However, writing rules to express a model that demonstrates the correlation between structural characteristics and biological function is still not an easy task. That is because there are very few structure-function correlations that have been clarified up till now, so general rules or explanations are still not available.

Finally, we learned that machine learning can be used to extract the characteristics of primary structure. We think it is particularly important that characteristics can be extracted for negative sets as well.

We believe that by incorporating this kind of inference structure in future integrated database systems, it will be possible to clarify structure-function relationships.

References

1. Kuhara, S., Satou, K., Takehara, H., Furuichi, E., Takagi, T., and Sakaki, Y., A Deductive Database System PACADE for the Three Dimensional Structure of Protein, In Proceedings of the Twenty-Fourth Annual Hawaii International Conference on System Sciences, Vol. 1, pp 653-659, January 1991.
2. Takagi, T., Satou, K., Kuhara, S. and Furukawa, T., Genomu Joho No Detabesuka Ni Mukete, Johoshori

Gakkai Detabesu Kininyoka To Wakushoppu [Toward a Database for Genome Data, Information Processing Society of Japan Database Gold Seal Course and Workshop], 47-56 (1991).

3. Kuhara, S., Satou K., Furuichi, E., Takiguchi, K. and Takagi T., Eneki Suiron Kino O Oyoshita Tampakushitsu Sanjikoze Kaiseki Detabesu Shisutemu PACADE, Dai 2 Kokai Wakushoppu "Hito Genomu Kaiseki To Johoshori Gijutsu" [Protein Tertiary Structure Analysis Database System PACADE Utilizing Deductive Logic Functions, 2nd Public Workshop "Human Genome Analysis and Data Processing Technology"], 162-165 (1991).

4. Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T., Negateibu Mochifu No Kikai Hakken, Dai 2 Kokai Wakushoppu "Hito Genomu Kaiseki To Johoshori Gijutsu" [Machine Discovery of Negative Motifs, 2nd Public Workshop "Human Genome Analysis and Data Processing Technology"], 62-65 (1991).

5. Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T., EFS No Gakushu Kanosei To Makutampakushitsu Ryoiki Yosoku E No Oyo, Johogaku Kiso [EFS Learning Capability and its Application in Predicting Transmembrane Protein Domains, Foundations of Data Processing] 23-1, 1-8 (1991).

6. Arikawa, S., Kuhara S., Miyano S., Shinohara, A., and Shinohara, T., Learning Algorithm for Elementary Formal Systems and Its Experiments on Identification of Transmembrane Domain, In Proceedings of the Twenty-Fifth Annual Hawaii International Conference on System Sciences, Vol 1, pp 675-684, January 1992

Restructuring of Sequence Data Using Profiles

92FE0879E Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 16-19

[Article by Osamu Goto, Saitama Cancer Center Laboratories]

[Text]

Background and Goals

It is thought that roughly 100,000 genes comprise the human genome, and we are trying to decode the sequences of all genes through the Human Genome Project. One of the most important topics is to learn the functions of those genes and their interrelationships. One basic method of doing so is to compare genes with the sequences of the proteins for which they code, and to compare both with those of other organisms. If sequence similarities can be found between several genes (or their products), we can surmise that these genes are descended from a common ancestor, and that they perform some kind of common function. Therefore, this provides us with powerful clues when we make analogies from data of known genes about the functions of other genes. In the same manner, it is useful to compare sequences to predict high level structures of proteins based on the known three dimensional structures of related proteins. Further, localized sites with especially high homology (conserved sites) indicate that

they were conserved during the course of evolution and suggests that they are functionally important locations.

When we actually compare sequences, mainly we search for similar sequences or line up related sequences (alignment). Sequence alignment means finding matching residue sequences in separate genes (or their products), and lining them up. The main goal of this research topic is to aid in predicting structure and function as discussed above by grouping and aligning bunches of genes or proteins that have mutually homologous sites, and by extracting the unique characteristics of those groups. We will express the extracted unique characteristics as true vectors called profiles, and construct a database with each profile as a single entry.

Details of Study

As mentioned above, the course of this research is divided into four steps. More specifically, (1) we will search out mutually related genes or proteins and group them. (2) We will align sequences within these groups. (3) We will extract the unique, shared characteristics of those groups from the aligned bunches of sequences. (4) We will construct a database with the unique characteristics expressed as profiles in separate entries. In the database, to carry out a more highly precise search, [word missing] will be fed back to (1), and to increase reliability, the characteristics extracted in (3) will be fed back to (2). Although the above four processes are mutually related, it is possible to consider them as independent research topics. To succeed in all four simultaneously is impossible at present, so we decided to focus on (2) and (3) to conduct our study. We chose the cytochrome P-450 multiple gene family as a model and conducted a study to see specifically how much useful knowledge the above method would yield.

Results

I. Improvement of Multiple Sequence Alignment by Repetition

The alignment of multiple sequences is in itself an extremely difficult problem, and it is necessary to devise methods to increase reliability and calculating efficiency. This year we proposed the following method and checked out its practicality.

1. Obtain preliminary alignment of multiple (N types of) sequences by an appropriate method.
2. Divide the sequences into two groups while maintaining their internal alignment. Arbitrarily choose one of the $2^{N-1}-1$ ways of dividing them. If there is a row consisting only of deletions within each group, delete that row.
3. Calculate the best alignment between two groups using dynamic programming methods.
4. Repeat 2 and 3 until the index expressing a "good" total alignment converges on a constant value.

The alignment obtained with this method is not necessarily the best in the narrow sense of the word, but from a practical standpoint it has high reliability. The run time for a protein sequence of conventional length of about $N \leq 10$ is not such a great burden compared with one of 2^N . Figure

1 shows that alignment is improved by repetition. (In the graph, the smaller the value on the vertical axis, the better the alignment.) The stairstep lines beginning on the upper left were generated when preliminary alignment was not performed, and the stairstep lines beginning on the lower left were generated after preliminary alignment was performed. Each line follows a course in which a different random number row was used. This demonstrates that the above method, including preliminary alignment, is effective in obtaining automatic multiple sequence alignment based on an objective standard in a relatively short time.

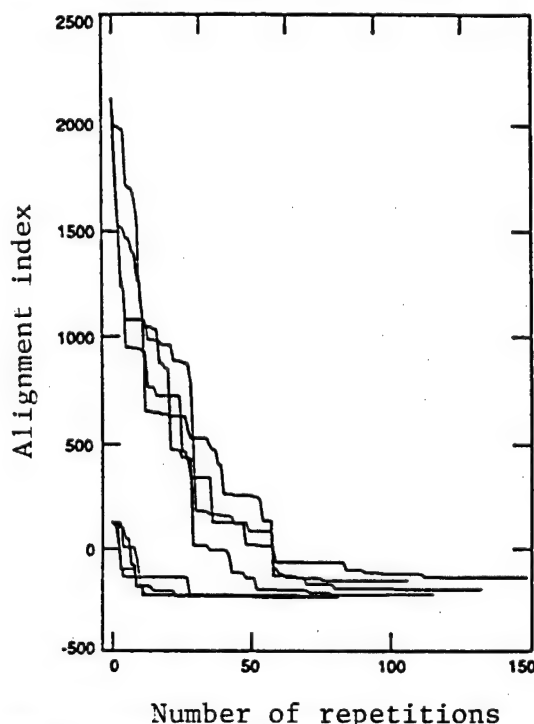


Figure 1. Improvement in Multiple Sequence Alignment by Repetition

II. Prediction of Cytochrome P-450 Protein Structure in Higher Animals

Cytochrome P-450 comes from a typical multiple gene family, and in the human genome it consists of 50 or more true genes. It has many physiological functions including the metabolism of various fatty acids such as steroid hormones, prostaglandins, and of external drugs and pollutants. Moreover, it is widely distributed in nature from bacteria to higher plants and animals, and at present amino acid sequences in nearly 200 species have been determined. We made a comparative study of these amino acid sequences. We demonstrated that all existing P-450 genes derive from a common ancestor, and that basically the high level structure of the protein is conserved. This year we performed a detailed comparison of the sequence of bacterial P-450 (P-450cam), the only one for which the three-dimension structure has been clarified, and the sequence of the drug metabolizing P-450 found in higher

animals, and attempted to identify the substrate recognition site of the drug metabolizing P-450. Less than 20% of the amino acid sequences of bacterial and drug metabolizing P-450 match, but as shown in Figure 2, profiles predicting hydrophobicity and secondary structure in both show extremely high homology except for one region. The three dimensional structure of P-450cam and the drug metabolizing P-450 substrate binding site predicted by the alignment in Figure 2 corresponded well with the results of genetic engineering experiments, and domains having abundant amino acid changes (Figure 3) show a generally good correspondence. These results show that the drug metabolizing P-450 has evolved so that it can adapt to the metabolism of a variety of substances, and at the same time, the amino acid sequence comparison used in the profile is a powerful method for predicting the structure and function of proteins.

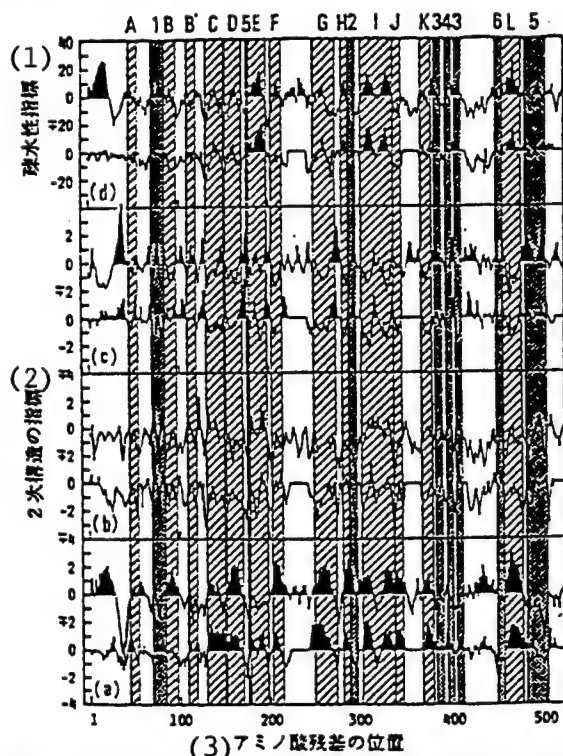


Figure 2. Comparison of Profiles of Bacterial and Drug Metabolizing P-450

Key: 1. Hydrophobicity index; 2. Secondary structure index; 3. Position of amino acid residues

Discussion

As shown in the example of cytochrome P-450, by using a profile it is possible to obtain reliable alignment between distantly related sequences in which only 20% or less of the amino acids match, and this suggests that it is applicable to a variety of prediction problems. Further, the method of

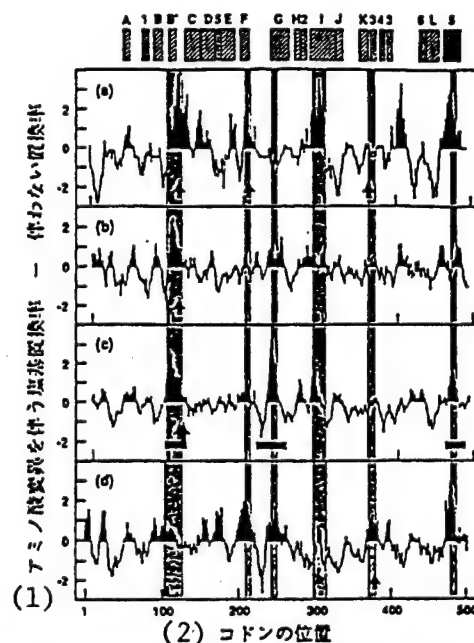


Figure 3. Difference in Ratios of Base Substitutions Accompanied by and Not Accompanied by Amino Acid Substitutions in Various Drug Metabolizing P-450 Sites
Key: 1. Ratio of base substitutions accompanied by amino acid substitution - ratio of base substitutions not accompanied by amino acid substitution; 2. Codon location

multiple sequence alignment by repetition that we proposed is still not perfected, and we want to add improvements to enable the alignment of even more sequences at an even higher speed in the future. The present human and physical organization is insufficient for implementing steps (1) and (4) of the four step process we described above. Enhancement of the organization, including joint research, is a topic we must address in the future.

References

1. Nishimoto, M., Gotoh, O., Okuda, K., and Noshiro, M., Structural analysis of the gene encoding rat cholesterol 7 alpha-hydroxylase, the key enzyme for bile acid biosynthesis, *J. BIOL. CHEM.*, 266, 6467-6471 (1991).
2. Kizawa, H., Tomura, D., Oda, M., Fukamizu, A., Hoshino, T., Gotoh, O., Yasui, T., and Shoun, H., Nucleotide sequence of the unique nitrate/nitrite-inducible cytochrome P-450 cDNA from *Fusarium oxysporum*, *J. BIOL. CHEM.*, 266, 10632-10637 (1991).
3. Kubota, M., Sogawa, K., Kaizu, Y., Sawaya, T., Watanabe, J., Kawajiri, K., Gotoh, O., and Fujii-Kuriyama, Y., Xenobiotic responsive element in the 5' region of the human P-450c gene., *J. BIOCHEM.*, 110, 232-236 (1991).
4. Gotoh, O., Substrate recognition sites in cytochrome P-450 family 2 (CYP2) proteins inferred from comparative analyses of amino acid and coding nucleotide sequences, *J. BIOL. CHEM.*, 267, 83-90 (1992).

**Application of an Object-Oriented Model for
Management of Hypothesis Knowledge Data**
*92FE0879F Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 20-22*

[Article by Teruo Koyama, R&D Dept., Science Information Center]

[Text]

1. Background and Goals

In the analysis of knowledge data in which the data structure is not known beforehand, as is the case with genetic data, one important topic is to generate hypotheses concerning the knowledge structure and then gradually clarify the structure of the knowledge through verifying these hypotheses. During the process of creating and verifying hypotheses, human beings are superb in their ability to create the hypothesis, but it is essential to fully utilize computer capabilities to verify the hypothesis when working with a massive database. My research for 1991 concerned the effectiveness of applying an object-oriented model in constructing an environment for the efficient creation and verification of hypotheses in a harmonious dialogue between computer and human being.

2. Genetic Data and Form of Hypothesis

Databases that presently deal with genetic data such as GenBank, EMBL, DDBJ, and the like store data in a format that permits mutual exchange. The contents of these databases vary, but what is important for the creation and verification of hypotheses concerning genetic data is the entry data, including the base or amino acid sequences as primary structures.

Typical hypotheses in handling genetic data make assumptions on the features of partial sequences of specific sequences or assume interrelations between partial sequences. For example, if a specific partial sequence of a protein is an α helix or a β sheet, we can consider this as typical hypothetical data. In the databases mentioned above, we can presume the data corresponding to the hypothesis is recorded in the FEATURE portion. In terms of managing this kind of hypothesis, problems include what kind of hypothesis it is, what is the reliability of the hypothesis or the circumstances under which it was created, and how do we define and preserve the process of manipulating and proving the hypothesis?

For the process of manipulating and proving the hypothesis, we can consider a series of procedures; for example, (a) assign a specific LOCUS to the partial string, (b) find the characteristic quantity of that partial sequence, and (c) determine the degree of certainty that the partial sequence has specific features. Moreover, in some cases, when we define a set of partial sequences that are assumed to have common characteristics, we also need an operation to extract the features that are characterized by that set and a procedure for determining the degree of homology with a certain partial sequence. In any event, it is necessary to extract the partial sequences of interest, calculate a characteristic value for the extracted partial sequence, and determine whether the partial sequence satisfies specific

conditions or not. These characteristic values and standards for making decisions must have a framework in which this kind of hypothesis data can be efficiently controlled because we can foresee many different things happening depending on the nature of the hypothesis in question.

Because various differences exist in the degree of certainty in the formation of a hypothesis, it must be linked to data that will act as a standard for judging its reliability. For this data we can consider not only data that directly describes the degree of certainty of the hypothesis, but also background data such as the creator (author) of the hypothesis, the literature in which the hypothesis was announced, the time the hypothesis was presented, and the existence of followup and countertheories.

In any event, hypotheses concerning genetic data must be managed by linking them to the entries in a genetic database. When we consider the large number of entries and hypotheses, and the sharing of data for maintaining a cooperative relationship among multiple researchers, it will be desirable for this to be managed under a single database management system in a form that integrates genetic database entries and hypothesis data.

Below I will discuss the results of a study on the applicability of an object-oriented model as the framework for this kind of hypothesis management with respect to genetic data.

3. Object-Oriented Model and Genetic Data

I believe that the advantages of using an object-oriented model for hypothesis management in genetic data are as follows:

1. A basic framework for manipulation of partial sequences can be shared.
2. A variety of data related to a hypothesis can be managed all together.
3. The method of calculating the characteristic values and the judgment standards can be defined in highly modular form corresponding to the nature of the hypothesis.
4. By using an object-oriented database, management of both database and hypotheses can be performed under a single environment.

In applying the object-oriented model as a framework for hypothesis management, it is desirable to manage the genetic database itself under the object-oriented model. By doing so, it will be possible to control the hypotheses as part of the database.

When we look at the data format of GenBank and the like, the sequence entry data consists of various fields. Basically, if we consider data elements as instance variables, we can map it directly onto the object-oriented model. We can think of REFERENCE and FEATURES as repetitive data with relatively complex structures, and if we consider each one as an object, it is possible to form links between the objects. As an actual problem, FEATURES is strongly related to the hypotheses, and we can think of it in the context of "hypotheses with relatively high reliability" or "actual examples of the hypotheses." Therefore, we can say that FEATURES is an object with the same basic structure as a hypothesis.

For a hypothesis expressed as an object, we must control many of the data and procedures mentioned above. Therefore, basically we can say that the characteristic values for each hypothesis and the various methods arise from the following structure:

- a. Hypothesis ID
- b. Details of hypothesis
- c. Reference entries
- d. Partial sequence assignments: start, end, limiting conditions
- e. Hypothesis reliability
- f. Author of hypothesis
- g. REFERENCE
- h. Various characteristic quantities

Procedures

- Assignment of combinations
- Methods for determining boundaries
- Calculation of characteristic quantities
- Verification of details of hypothesis
- Verification of reliability.

I studied the methods of realizing a hypothesis management framework under this kind of setup. At present, typical environments that support object-oriented programming include:

- C++
- Smalltalk 80
- Common Lisp Object System (CLOS).

Among these, C++ is characterized by good running efficiency, and Smalltalk 80 features an abundance of basic object/methods and a user friendly programming environment. In a sense, CLOS offers the most powerful object-oriented programming environment, and it supports multiple accession and dynamic class change. Moreover, CLOS is characterized by its ease in linking various knowledge data programs. On the other hand, from the aspect of linking with object-oriented databases, both C++ and Smalltalk 80 have interfaces with several existing object-oriented database management systems, but it will be difficult to link CLOS with an existing DBMS.

Looking at the above conditions as a whole, we can say that the best combination for a programming environment will be C++ and the object-oriented database management system. At present I am trying to set up a system under this kind of environment.

Creation of Integrated Database Containing Protein Sequence and Knowledge Data

92FE0879G Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 23-25

[Article by Yasuhiko Seto, Seiji Isoyama, and Yoshinori Takeuchi, Protein Research Promotion Society (non-profit corporation)]

[Text]

Background and Goals

Since 1975, when we began creating peptide and protein literature databases, we have constructed the following protein-related databases (DB). These have become the basic data for analyzing the appearance of a protein on the molecular level.

Database Title: Example of Entry Details

- (A) LITDB (LITERATURE-DB): lit. no./reference/author/keyword, keyphrase
- (B) SEQDB (SEQUENCE-DB): lit. no./protein/organism/sequence/comment
- (C) FRAP LIBRARY: lit. no./protein/fragment peptide (motif)/comment
- (D) KNOWLEDGE-DB: lit. no./protein/ligand/sequence of functional site
- (E) MUTDB (MUTANT-DB): lit. no./protein/mutation/function, stability
- (F) STRUCTURE-DB: lit. no./protein/secondary structure element

For proteins, sequence data is the most basic kind. We compiled and organized data from scholarly journals, and found that more than 90% of sequences are now determined from DNA sequences. In the past, first we had a protein with chemical and biological characteristics which was refined and sequenced, but progress in genetic engineering has enabled us to easily determine a protein's sequence from DNA without knowing its characteristics. This has brought about two problems. First, which sequence should we begin with in a huge gene? On the other hand, what kind of protein does that specific sequence code for? One goal in creating the integrated database described above was to answer these questions from the standpoint of the protein. Our goal is first to collect the massive wealth of information that is scattered about over a wide area, and then create a database that many kinds of researchers, including experimental researchers, can easily use.

Details of the Integrated Database

(A) Literature Database: LITDB

This database surveys roughly 1,000 journals for protein references presented on the molecular level. Among the total of 174,644 input reference items in 1991, 9,949 were related to sequence. The following keywords were assigned to references containing experimental and knowledge data: 1. Sequence determination (seq determination 4332 ('91)). 2. Sequence comparison (seq comparison 3016 ('91)). 3. Mutant sequences (mutant 1468 ('91)). 4. Interactive sites (binding site 1783 ('91)). 5. Three dimensional map (stereo structure 330 ('90)). These references were used in the creation of the following databases: 1. SEQDB, 2. FRAP LIBRARY, 3. MUTDB, and 4. KNOWLEDGE-DB. In databases other than the LITDB, multiple bibliographic entries are eliminated by relational reference back to the LITDB.

(B) Amino Acid Sequence Database: SEQDB

For the SEQDB we extract data from the scientific journals surveyed in the process of creating the LITDB, and input them directly. The number of inputs for 1991 were 6,354, and as of this writing in March 1992, we have stored a total of 33,112 entries for a total of 10,702,314 amino acid residues. The details of the data include the protein name, organism, sequence, and a minimum of comments. Details of data concerning sequence are referenced in the LITDB. This SEQDB was transferred to SYBASE. In this process we organized the contents of several entries. First, we standardized the scientific names of the organism from which the protein was obtained. For those organisms where the scientific name was unclear, we recorded the name of the lowest known taxon that includes that organism as determined from the common name of the organism. For each entry we recorded whether the amino acid sequence was determined from mRNA, gene, or protein. We created a taxonomic table of organisms, and this enables an arbitrary search of proteins by species, genus, family, and so on.

(C) Fragment Peptide Library: FRAP LIBRARY

This library contains experimentally determined biological activity, fragment sequence binding activity and fragment sequences (motifs) conserved among homologous proteins. This year we put our efforts into organizing motifs. We extracted motifs from sequence comparison data using sequence determination references that we surveyed in the process of creating the sequence database. The motifs were extracted according to the following three standards. 1. A length of 5 to 20 residues. 2. Roughly three motifs to a single protein. 3. Proteins with multiple domains are referenced by each domain. FRAP has a total of about 10,153 stored sequences, including 9,854 conserved sequences. The data for 1991 was extracted from 3,016 references. The lengths of the motifs and the number of entries (in parenthesis) were: 5 (2,077), 6 (2,406), 7 (1,646), 8 (870), 9 (318), 10 (113), 11-20 (110). When we performed sequence comparisons between proteins with weakly homology, we were easily able to extract conserved sequences using our knowledge of chemistry and biology.

(D) Knowledge Database

When a protein expresses a function, a specific site or sequence plays an important role. References containing this kind of knowledge can be extracted from the LITDB by using several keywords. This year we collected, organized and made a database of references concerning sites and sequences that interact with ligands and were extracted with the keyword "Binding." This kind of knowledge is often used for site directed mutagenesis in genetic engineering, so it was necessary to organize and integrate the references with the mutant database. The number of references we processed were 71 (1979), 89 (1980), 107 (1981), 123 (1982), 116 (1983), 263 (1984), 379 (1985), 582 (1986), 913 (1987), 1,143 (1988), 1,456 (1989), 1,520 (1990), and 1,783 (1991). The data format was: Reference number, protein name, ligand, functional site, sequence/comments. In the sequence column, we recorded data for comparison with other database sequences.

(E) MUTDB: Mutant Database

We surveyed details of references searched from LITDB with the keywords "Mutant, Mutagenesis, Mutation, Variant, Substitution, Replacement," and made them into a database. Data entries contain the details of sequence mutation, characteristics of the changes, sequence of the five residues of the N terminus, and mutual references to other databases. This year we studied problems in transplanting this to a relational database, and we transferred part of the items to tables. The creation of this mutant sequence database was performed jointly with the Protein Engineering Institute.

(F) Three Dimensional Structure Database

We collected data corresponding to sequences and secondary structure elements from references in which X-ray analysis of proteins was reported, and three dimensional structures were published. The number of references we used in data collection were: 177 (1985), 222 (1986), 261 (1987), 282 (1988), 362 (1989), 330 (1990), 414 (1991) for X-ray analysis and 132 (1985), 188 (1986), 262 (1987), 251 (1988), 259 (1989), 255 (1990) and 391 (1991) for three dimensional structures.

Discussion

In traditional text type databases, the notation for each item recording the data content was in a chapter format, and even if the various items did not always stand by themselves, it was possible to look through them and use the database. In transferring data to SYBASE, it was necessary to record each item of data precisely in tables and in a uniform format. The transfer took much longer than we originally expected. For example, making tables in a form in which the journal title, volume, number, page, protein name, and its origin were all coordinated was a very laborious task. There were problems in the recognition of almost every scholarly journal and author who submitted a manuscript.

The technique of genetic engineering is gradually expanding its territory into chemistry, biology, pharmacology, medicine and the like, and in the sequence database there are more than 400 scholarly journals recording determined sequences. The collection of data for the integrated utilization of protein and nucleic acid sequences, various experiments that involve the sequences, and knowledge data is becoming increasingly difficult. Input of the sequences themselves can be done on line, but in the future, we must sufficiently study the formats and methods for inputting knowledge. The importance of this becomes clear when we consider that the online input of data into databases will become commonplace with greater expansion of computer networks.

On the other hand, there is a major problem because data input into computer systems via the networks is intractable. As one example, among the 4,332 (1991) sequence determination references, there were 2,323 with recorded accession numbers. This is not always because people do not understand about on line input, for among the journals we find writers who say they cannot go along with the views of the editors, that the information will appear soon

in a database, that they are dissatisfied with the organization that creates databases, that accepting data that has not undergone the peer review it does in scholarly journals is meaningless, and the like.

The databases we described, are still weak entities that search while making inquiries by line input. However, in the near future, we expect to have a full screen user interface through the help of Masami Hagiya's group.

We who are compiling a variety of protein sequence-related data from scholarly journals must solve problems such as standardization of data, user interface, on line input and network utilization in the process of creating an integrated database.

Organization of Genome Physical Map Data

92FE0879H Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 p 26

[Article by Eichi Soeda and Xiao-Ren Tang, Gene Bank, Institute of Physical and Chemical Research]

[Text]

Background and Goals

The goal of this research is to study and set parameters for creating a format for physical maps for DNA clones, especially aligned clone maps, that will be produced during the development of the Genome Project. Further, during the creation, we will complete the format for a map for STS (sequence tagged sites) that are produced.

Details of Study

We studied mass use of STS for chromosome markers.

Results

1) We tried cycle sequencing and found it was possible to directly sequence double stranded DNA.

2) Using an automatic DNA purifier we established a mass preparation method for pUC vector series chromosome markers and performed direct cycle sequencing. As a result it was possible to read a minimum of 300 bases, and this yielded sufficient data for STS design. Preparation of 320 template DNAs is possible in 1 day.

Discussion

As genome research progresses, the number of chromosome markers is increasing geometrically. In order to respond to this, we used an automatic DNA purifier (Clavo [phonetic] p100) for mass preparation of pUC vector series markers and performed cycle sequencing. As a result, production of a large amount of STS sequences was possible, but the setting of the remaining oligonucleotide and PCR conditions are topics we must address next year. Concerning the construction of an STS database, we will introduce the Johns Hopkins University Genome Database (GDB) after the establishment of a common format that can be used throughout the world.

References

1. Ozawa, K., Sudo, T., Soeda, E., Yoshida, M. C., and Ishii, S., Assignment of the Human CREB2 (CRE-BP1) Gene to 2q32, GENOMICS, 10, 1103-1104 (1991).
2. Furuno, N., Nakagawa, K., Eguchi, U., Ohtubo, M., Nishimoto, T., and Soeda, E., Complete Nucleotide Sequence of the Human RCC2 Gene Involved in Coupling Between DNA Replication and Mitosis, GENOMICS, 11, 459-461 (1991).
3. Ozawa, K., Murakami, Y., Eki, T., Soeda, E., and Yokoyama, K., Mapping of the Gene Family for Human Heat Shock Protein 90 α to Chromosomes 1, 4, 11 and 14, GENOMICS, 12, 214-220 (1991).
4. Tashiro, H., Ozawa, K., Xiaoren Tang, Nikai, H., Eki, T., Murakami, Y., Soeda, E., and Yokoyama, K.; Single DNA Marker Generated by "YAC-Alu PCR" That is End-Specific; JPN. J. HUMAN GENETICS, 36, 229-243 (1991).
5. Euy Kyun Shin, Matsuda, F., Nagaoka, H., Fukita, Y., Imai, T., Yokoyama, K., Soeda, E., and Honjo, T., The EMBO Journal, 10 (12), 3641-3645 (1991).

Study on High-Level Searching of Protein Three-Dimensional Structural Data Using Deductive Database Methodology

92FE0879I Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 27-30

[Article by Toshihisa Takagi (presently at Human Genome Analysis Center, Dept. of Medicine Laboratory, University of Tokyo), and Kenji Sato, Data Processing Educational Center, Kyushu University]

[Text]

Background and Goals

As genome research advances, it is necessary to store and search mass amounts of data on a variety of different levels such as base sequence and amino acid sequence data, nucleic acid and protein three dimensional structures and the like. If all we need to do is merely accumulate a large amount of this data, it can be accomplished using existing relational database methods. However, if we want to perform a complex, high level search to clarify the relationship between the three dimensional structure of a protein and its function, for example, the search functions on existing databases are inadequate. In the field of computer science, R&D on deductive database methods has made rapid progress in recent years, and we approached this problem using these methods.

A deductive database system is a database system that integrates logic programming techniques and relational database techniques. The deductive database is characterized by the fact that it has a powerful inference function and can efficiently perform complex, high level searches. In the past, there was the case in which a system integrating PROLOG, which is a representative logic programming language, and a relational database was used for genome analysis, but there is a problem with PROLOG because to

assure that the inference function will cease and to improve efficiency, the user must limit the inference. Therefore, methods have been studied that require no inference limitations but still efficiently process large amounts of data, and several processing methods including the magic set method have been proposed. Using these latest methods, we have succeeded in performing an efficient search of high-level engineering data with complex structures on a large scale.

In this research we will determine inference functions required for a high level search of protein three dimensional data and develop a question processing method for a deductive database that performs efficiently. In this case for the question processing method we will use the magic set method as a basis rather than the PROLOG processing method. Next we will create a system based on this method, evaluate it, and establish a technique for efficient searching of protein three dimensional data.

Details of Study

Two major research topics emerge in the analysis of genetic data. One is the clarification of the identification of genes on DNA and the manner in which their expression is regulated. The second is the clarification of the functions of the proteins coded by the genes. Along with advances in methods for determining gene base sequences, a large number of sequences have been determined and their databases are under construction. At the same time, now that the amino acid primary structure of proteins is being determined from base sequences, if we can clarify the relationship between protein structure and function, we will be able to estimate the function of a protein by determining the gene base sequence. To clarify this relationship between protein structure and function, analysis that includes not only primary structure but also secondary and tertiary structure data is indispensable. Therefore, using the protein tertiary structure database Protein Data Bank (PDB), we constructed a database system PACADE (Protein Atomic Coordinate Analyzer with Deductive Engine) equipped with deductive functions from rules for protein high level structural analysis.

Almost all database systems for protein tertiary structure analysis constructed up till now employ a relational database management system and SQL as the question language. A database has also been constructed that applies logic programming in searches and uses PDB data as PROLOG "facts" instead of using SQL. Further, recently an object-oriented database system has been constructed that incorporates a protein structure hierarchy. The deductive database system PACADE that we developed is based on a method called the magic set method, and it displays authority especially in searches that use recursive relationships. By expressing protein characteristics in rules with

PACADE, it is possible to abstract protein secondary and tertiary structures. Therefore, we believe that this system will become a powerful means for clarifying the basic structure-function relationships of proteins in the future.

Results

If we begin with our conclusion, at present we have learned that searches for hydrophobic clusters, dipole moment, repeating structures, barrel structures and the like can be performed easily and efficiently by using the PACADE system that we developed.

We originally developed PACADE with the goal of performing high level searches for protein tertiary structure, but thereafter we added data on secondary structure, and at present, it is possible to search for secondary, super-secondary, and tertiary structures. This system was constructed on a SUN workstation. This system consists of three types of files (PDB file, Database, and Rule file) and three modules (SYBASE, Extractor, and Deductive Engine).

The PDB file has data in general text form such as three dimensional coordinate values and the like. The Database is a relational database created by the Extractor from the PDB file, and it contains data such as three dimensional coordinate values, distances between residues, secondary structures, and relations among secondary structures (for example, parallel, antiparallel, etc.) in 17 relational tables. The Rule File contains rules written in text format, and these are evaluated in the Deductive Engine. SYBASE is a commercially available relational database management system. It performs database management, accepts commands from the Deductive Engine, performs the search and returns the results. The Extractor converts from the PDB file into SYBASE input format. The Deductive Engine consists of two submodules.

(1) Rule Transformer: When this module receives a question from the user, it reads rules from the Rule File and converts them into rules for an efficient search using the magic set method.

(2) Bottom-up Evaluator: When this module receives a question from the Rule Transformer, it evaluates it using the semi-naive method and returns an answer. Before performing the evaluation it must search SYBASE and create a fact.

Below is shown an example of search rules (partial) for a Greek-key structure as an example of a super-secondary structure search on PACADE.

We applied these rules and performed searches for super-secondary structures on the proteins thermolysin, immunoglobulin, and prealbumin. As a result we learned that for the Greek-key structure in each protein, we had to expand the definition of parallel to an angle of 70°.

```

meander_n(A, [B, C], 3, P) :- hairpin(A, B, P), hairpin(B, C, P).
meander_n(A, [B|L], N1, P) :- hairpin(A, B, P), meander_n(B, L, N, P), N1 = N + 1.
greek_even_r(A, [], D, 2, P) :- hairpin(A, D, P).
greek_even_r(A, L, D, N1, P) :- not_coils(A, B, P), neighbour(A, D, P),
                                double_anti_parallel(A, D, P),
                                greek_odd(B, L1, D, N, P),
                                append([B], L1, L), N1 = N + 1.
greek_even_l(A, [], D, 2, P) :- hairpin(A, D, P).
greek_even_l(A, L, D, N1, P) :- not_coils(C, D, P), neighbour(A, D, P),
                                double_anti_parallel(A, D, P),
                                greek_odd(A, L1, C, N, P),
                                append(L1, [C], L), N1 = N + 1.
greek_odd(A, L, D, N1, P) :- greek_even_r(A, L1, B, N, P),
                             greek_even_l(C, L2, D, N, P),
                             append(L1, [B], L), append([C], L2, L3),
                             L = L3, N1 = N + 1.
hairpin(A, B, P) :- not_coils(A, B, P), neighbour(A, B, P),
                   double_anti_parallel(A, B, P).

```

Discussion

As noted above, we confirmed to a certain extent that the PACADE system we developed is effective in searches for complex protein structures. Because PACADE is equipped with inference functions, we hope that it will be easy to create and verify biological hypotheses by using this system. In the future, we plan to add topological data on secondary structure to the database so that high level searches can be performed even more easily. Furthermore, we have plans to improve the user interface so that biological researchers can use this system easily.

Further, although not directly related to this research theme, we performed the following research this year on the genome database. We will report the details of this research separately.

Development of ODS: Overlapping Oligonucleotide Database

We divided continuous base sequences into overlapping oligonucleotides (length 8 bp), and together with data concerning locations of functional sites, constructed a deductive database. We used data from GenBank for this data. This database can be used to predict promoter regions and the like.

Survey of Present Status of Databases for Genome Analysis and Problems

Creating various kinds of databases for genome analysis will be indispensable. However, genome data has many

properties that differ from the business and engineering data handled by current databases, and it will be impossible to deal with genome data using only the technology of present databases. Therefore, we surveyed the present status of databases used for genome analysis with the goal of developing new database technology for genome analysis and clarified the problems involved.

Development of Editing Software for GenBank Data

The data distributed in GenBank contains omissions and duplications, which makes it difficult to use for genome analysis in its present state. For example, the base sequence of a single gene is divided and stored as multiple entries. Therefore, while taking into account their biological significance, we are presently developing software to reconstruct entries by gene, by chromosome, by species and the like, and we are nearly finished with software that classifies the entries by gene.

References

1. Takagi, T., Satou, K., Kuhara, S., and Furukawa, T., Genomu Joho No Detabesuka Ni Mukete, Johoshori Gakkai Detabesu Kininyoka To Wakushoppu [Toward a Database for Genome Data, Information Processing Society of Japan Database Gold Seal Course and Workshop], 47-56 (1991).
2. Nishikawa, A., Sakamoto, N., Takagi, T., Sakaki, Y., and Ushijima, K., Idenshi Deta No Bunruiho Ni Tsuite,

Dai 2 Kokai Wakushoppu "Hito Genomu Kaiseki To Johoshori Gijutsu" [Classification of Genetic Data, 2nd Public Workshop "Human Genome Analysis and Data Processing Technology"], 78-81 (1991).

3. Kuhara, S., Satou, K., Furuichi, E., Takiguchi, K., and Takagi, T., Eneki Suiroon Kino O Oyoshita Tampakushitsu Sanjikoze Kaiseki Detabesu Shisutemu PACADE, Dai 2 Kokai Wakushoppu "Hito Genomu Kaiseki To Johoshori Gijutsu" [Protein Tertiary Structure Analysis Database System PACADE Utilizing Deductive Logic Functions, 2nd Public Workshop "Human Genome Analysis and Data Processing Technology"], 162-165 (1991).

4. Sakamoto, N., Takagi, T., Sato, K., and Sakaki, Y., Development of Overlapping Oligonucleotide Database and Its Application to Searching for Signal Sequences Over the Human Genome, 1992 Nen Johogaku Shimpoijium [1992 Information Science Symposium], 37-46 (1992).

Research on Determining Amino Acid Sequence Characteristics by Chemical Structure of Ligands
92FE0879J Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 31-34

[Article by Takaaki Nishioka and Mikita Suyama, Chemical Institute, Kyoto University]

[Text]

Background and Goals

To predict the biological significance of coded base sequences in the genome, and more specifically, to predict the functions performed by proteins in the body from the protein amino acid sequences that are products of the genome, we study their homology with the amino acid sequences of proteins whose functions have already been determined. In doing so, we can assume that protein is the same as a known protein if it shows homology over the whole amino acid sequence from amino terminus to carboxyl terminus, or that it belongs to a homologous family if it shows homology over a long region corresponding to a domain. At present, however, when the sequence whose function we are trying to predict shows only localized homology with sequences of several different known proteins, we find it extremely difficult to predict its function. It is thought that the human genome contains many cases in which this kind of localized homology appears. If a dictionary were to be created that links function to localized sequences consisting of about 20 or 30 amino acid residues, this dictionary could be used to predict the functions of proteins coded in the human genome. It will be necessary to select sequences with functions that exist universally in as many proteins as possible and that preserve their biological significance even when the protein is divided up. In this research theme, we focused on the fact that almost all functional proteins distinguish specific compounds (ligands) from all other substances within the body and they exhibit binding specificity, and that the chemical structures of ligands are diverse and we can divide up their chemical structures into even smaller partial chemical structures. Our goal was to create a dictionary relating the functions of short amino acid

sequences to the partial chemical structures of ligands and utilize this in genome analysis. In 1991 we constructed the database needed for creating the dictionary.

Details of Study

To create a dictionary linking the amino acid sequences of proteins and the chemical structures of ligands, it is necessary to construct a database linking the chemical structures of ligands that recognized by proteins with the amino acid sequences of proteins whose sequences have already been determined. Among proteins there are enzymes that have been clearly linked with ligands, so we decided to construct a database linking enzyme amino acid sequences with chemical structures (Enzyme-Reaction Database).

To readily distinguish between proteins, we used the EC coding of the IUB [International Union of Biochemistry]. We linked amino acid sequences and proteins by NBRF-PIR entry number. For enzyme protein ligands we used enzyme reaction substrates, reaction products, coenzymes, inhibitors, activators and the like. To express the ligands, we require binding tables noting the name of the compounds, their structures and the like, but in this database we chose to list only the names of the compounds. We recorded the binding tables together with the names of the compounds on the MACCS chemical data management system (MDL Co., USA) installed on the Chemical Institute's FACOM M780, and made it a separate construction. That is because a chemical data management system capable of running on the UNIX had not been developed when the institute was started.

The original database was created in flat files, and we updated it along with the NBRF database using a program written in FORTRAN. Then we decided to transfer it to a relational database for use in various kinds of searches and motif extraction.

Results

(1) Enzyme-Reaction Database

This database consists of (1) enzyme names (including common names), EC numbers, types of catalyzed reactions, and chemical reaction formulae, (2) names of ligand compounds, (3) enzyme-linked illnesses and names of inherited diseases and MIM numbers based on McKusick classification, (4) entry numbers on the NBRF-PIR amino acid sequence database, and (5) comments. Presently in November 1991, we have 1,103 types of enzymes recorded, 6,620 NBRF-PIR entry numbers, names of 1,756 ligands, and 182 names of enzyme-linked illnesses.

In other words, among the total of 2,477 types of enzymes recorded and classified by the IUB, amino acid sequences have been determined for roughly 45% and recorded in NBRF-PIR.

Among the ligands are recorded not only compounds synthesized in the body by metabolism but also chemical substances synthesized by organic chemists as medicines and pesticides. Of the ligands, the chemical structures of 842 compounds are recorded in the MACCS system, and it is possible to search by chemical structure or partial chemical structure. In other words, if a chemical structural

formula is input, the name of the compound, including its chemical structure, is output. For example, if we input chemical structures of adenosine, the output is not only adenosine, but also AMP, ADP, ATP, Coenzyme A, NAD, NADP, FAD, RNA, etc.

For the names of enzyme-linked illnesses, we included not only hereditary illnesses caused by a deletion or mutation in the enzyme gene, but also all the names of illnesses suspected to be related to that enzyme.

(2) Transfer to Related Databases

For performing high level searches among data items, we first transferred the data to SYBASE for decentralized process searching. We created seven types of tables corresponding to the tables for each item in the database, EC number-enzyme name, EC number-substrate name, EC number-inhibiting compounds, etc. Using this, when we searched with the statement "Display EC number and NBRF-PIR entry number for enzymes producing glutathione" in SQL, we got an output of three types of enzyme names, glutathione reductase (EC 1.6.4.2), lactoyl-glutathione lyase (EC 4.4.1.5) and glutathione synthase (EC 6.3.2.3), and their six types of NBRF-PIR entry numbers. Thus, we could link the name of the ligand output from the MACCS system with its related enzymes and their amino acid sequences.

(3) Extraction of Amino Acid Sequences Characterizing Ligand Structure

In the example shown in (2) above, the features of amino acid sequences recognizing the chemical structure of glutathione ought to be extracted as a common motif of the six types of amino acid sequences output in the SYBASE search. We tried multiple alignment of these six types of amino acid sequences, but we could not detect any common sequence segments. Then we performed multiple alignment for each enzyme and extracted sequence motifs, and by comparing the differences among the enzymes, we extracted the sequence motif characterizing chemical structure of the ligand.

(4) Extraction of Sequence Motif Characterizing an Enzyme

We decided to perform multiple alignment of the amino acid sequences for each enzyme and extract sequence motifs. For multiple alignment we decided to calculate the homology score for all possible pairs of sequences. First we paired the most similar sequences, then found the sequence most similar to the first two and so on in a recursive operation. In doing so we permitted branches in the nodes. We extracted the entry numbers from the database, extracted sequences from NBRF-PIR, and ran multiple alignment (a C language program) recorded in PERL language.

When we performed multiple alignment on a SUN SPARC station 2 for 4,600 sequences from enzymes in which two or more sequences (excluding fragments) have been determined, it took 37 CPU hours. The 211 sequences of the (H⁺)-transporting ATP synthase was the enzyme with the largest number of sequences, and alignment of these alone

took 3.35 CPU hours. We called the conserved residues discovered in this manner conserved sequences, and expressed the perfectly matching conserved amino acid residues with symbols other than letters when their physical properties such as electric charge, hydrogen binding capability, hydrophobicity, etc., were conserved. We divided the conserved sequences into five-letter motifs, and entered each in the dictionary as a characteristic motif.

Discussion

(1) Features of Enzyme-Reaction Database

Until now, no database has been constructed with the goal of studying the relation between amino acid sequence and the chemical structure of the ligand recognized by a protein. In this research we constructed such a database and transferred the data to a relational database, which enabled comprehensive and systematic extraction of the relevant protein amino acid sequences when searching by the name of a chemical substance. In this manner, we completed a database that will serve as a foundation for characterizing amino acid sequences by the structures of their ligands. Further, because it gives data concerning enzyme-related illnesses, we hope that it will benefit the Human Genome Project and therapeutic drug design.

(2) Search by Chemical Structure

At present, the data for the chemical structures of ligands (binding tables) is managed by the MACCS system on the FACOM OS MSP/OS4. Therefore, the disadvantage is that it is isolated from the relational database on the UNIX. Just recently, the DAYLIGHT system which is the first chemical structure data management system to run on the UNIX was developed by Weininger, et al. We are presently studying transferring this to the SUN workstation. With this system, because structural formulae are handled and expressed in a sentence type line-notation called SMILE, it will be possible to record chemical structures written in SMILE on the database.

(3) Sequence Motif Extraction

We performed multiple alignment of the 1,103 types of enzymes. Although some enzymes were classified as the same by enzymological criteria, we found that we could not extract sequence motifs well in cases where there is absolutely no sequence homology due to convergent evolution in different species and in cases where multiple, different domains are coded by a simple sequence in multifunctional enzymes. Therefore, after dividing the various enzymes beforehand into groups based on their sequence homology, we performed multiple alignment on each group and were able to extract motifs cleanly. At present, we are at the stage of adding studies on methods of storing, comparing, and characterizing motifs obtained in this manner.

References

1. Suyama, M., Ogiwara, A., Nishioka, T., and Oda, J., Construction and application of Enzyme-Reaction Database. Submitted for publication.

2. Suyama, M., Ogiwara, A., Nishioka, T., and Oda, J., Koso No Kino To Hairitsu Ni Kansuru Detabesu No Kochiku To Riyo, Dai 2 Kai Genomu Joho Wakushoppu [Construction and Use of a Database Concerning Enzyme Function and Sequence, 2nd Genome Data Workshop], Tokyo (1991).

3. Sumi, K., Nishioka, T., and Oda, J., Similarity graphing and enzyme-reaction database: methods to detect sequence regions of importance for recognition of chemical structures, *Prot. Eng.*, 4, 413-420 (1991).

Evaluation of Parallel Inference Machine in Sequence Analysis and Biological Databases
92FE0879K Tokyo MOE KEY AREA RESEARCH in Japanese Mar 92 pp 35-36

[Article by Katsuki Nitta, Kazumasa Yokota, Mikihiro Ishikawa, Makoto Hirosawa, Masanori Hoshida, Hidetoshi Tanaka, Noriyuki Todani, and Kentaro Onizuka, New Generation Computer Technology Development Agency (non-profit corporation)]

[Text] To collect and utilize data on nucleic acids and proteins, sequence analysis technology and database technology are needed. Both require knowledge processing techniques and database techniques as well as extremely powerful computers in order to realize high-level functions. The purpose of this research is to show that a parallel inference machine (PIM) and a deductive object-oriented database language running on it can be effective in sequence analysis and the construction of biological databases.

This year we worked with simulated annealing, which is the elemental technology behind multiple alignment, three dimensional dynamic programming, and tournament methodology, and we integrated them in a system called MASCOT (Multiple Alignment System in ICOT). In the databases, we concentrated on a trial graphic display of protein characteristics for user interface integration, a study of the application of deductive object-oriented concepts for knowledge base integration, and functional notation for proteins.

As a result of the above studies, we found that in sequence analysis we could achieve relatively good alignment with parallel processing in a practical amount of time even on sequences with low homology. In databases, we found that descriptions of enzyme reactions and data searches using those descriptions were easily realized with deductive object-oriented language.

To perform high level sequence analysis, it is necessary to make better use of our knowledge of organisms. A future theme will be the development of a system incorporating knowledge of organisms using knowledge processing techniques. Further, it is necessary to proceed with description experiments to see if deductive object-oriented language can be applied to express various phenomena, functions, and structures other than enzyme reactions. Among the functions required for the creation of an integrated database will be the need to apply data search technology to

expand existing database management systems for sequence homology searches.

References

1. Ishikawa, Hoshida, Hirosawa, Todani, Onizuki, Nitta, and Kanehisa: Heiretsu Suiron Mashin O Mochiita Tampakushitsu No Hairitsu Bunseki, Johoshori Gakkai, Johogaku Kiso Kenkyukai Hokoku [Protein Analysis Using a Parallel Inference Machine, Information Processing Society of Japan, Data Science Basic Research Report] 23-2, 1991.

2. Todani, Hoshida, Ishikawa, and Nitta: Heiretsu 3 Jigen Dainamikku Puroguramingu Ho Ni Yoru Tampaku No Hairitsu Kaiseki, Johoshori Gakkai, Dai 5 Kai Puroguramingu Kenkyukai Hokoku [Protein Sequence Analysis by Parallel Three Dimensional Dynamic Programming Methods, Information Processing Society of Japan, 5th Programming Forum Report], 1991.

3. Hoshida, Ishikawa, and Hirosawa: 3 Uei DP Majiho O Mochiita Maruchipuru Araitentu, Nihon Seibutsu Butsuri Gakkai Dai 29 Kai Nenkai [Multiple Alignment Using Three-Way DP Merging Methods, Japan Biophysics Society, 29th Annual Conference], 1991.

4. Ishikawa, Todani, Hoshida, Nitta, and Kanehisa: Heiretsu Shimyureddo Aniringu O Mochiita Marchipuru Araitentu, Nihon Seibutsu Butsuri Gakkai Dai 29 Kai Nenkai [Multiple Alignment Using Parallel Simulated Annealing, Japan Biophysics Association, 29th Annual Conference], 1991.

5. Hirosawa, Hoshida, and Ishikawa: Hairitsukankyori Kaiseki O Koryoshita Maruchipuru Araitentu Shisutemu, Nihon Seibutsu Butsuri Gakkai Dai 29 Kai Nenkai [Multiple Alignment System Incorporating Analysis of Distance Between Sequences, Japan Biophysics Association, 29th Annual Conference], 1991.

6. Ishikawa, Hoshida, Hirosawa, Todani, and Nitta: 3 Jigen Dainamikku Puroguramingu O Mochiita Tampakushitsu No Hairitsukaiseki, Johoshori Gakkai Dai 43 Kai Zenkoku Taikai Rombunshu [Protein Sequence Analysis Using Three Dimensional Dynamic Programming, Information Processing Society of Japan, Proceedings of the 43rd National Conference], 1991.

7. Todani, Hoshida, Ishikawa, Nitta, and Kanehisa: Heiretsu Shimyureddo Aniringu O Mochiita Maruchipuru Araitentu, Johoshori Gakkai Dai 43 Kai Zen Koku Taikai Rombunshu [Multiple Alignment Using Parallel Simulated Annealing, Information Processing Society of Japan, Proceedings of the 43rd National Conference], 1991.

8. Hirosawa, Hoshida, and Ishikawa: Tampakushitsu Hairitsukankyori Kaiseki O Mochiita Tampakushitsu No Sodosei Kaiseki Shisutemu, Johoshori Gakkai Dai 43 Kai Zen Koku Taikai Rombunshu [Protein Homology Analysis System Using Analysis of Distance Between Protein Sequences, Information Processing Society of Japan, Proceedings of the 43rd National Conference], 1991.

9. Ishikawa, Todani, Hoshida, Nitta, Ogiwara, and Kanehisa: Heiretsu Shimyureddo Aniringu O Mochiita Maruchipuru Araimento, Dai 2 Kai Kokai Wakushoppu "Hito Genomu Keikaku To Joho Kaiseki Gijutsu" Rombunshu [Multiple Alignment Using Parallel Simulated Annealing, 2nd Public Workshop "Human Genome Project and Data Analysis Technology" Proceedings], 1991.

10. Hirosawa, Hoshida, Ishikawa, and Todani: 3 Jigen Dainamikku Puroguramingu O Katsuyoushita Tampakushitsu No Araimento Shisutemu, Dai 2 Kai Kokai Wakushoppu "Hito Genomu Keikaku To Joho Kaiseki Gijutsu" Rombunshu [Protein Alignment System Utilizing Three Dimensional Dynamic Programming, 2nd Public Workshop "Human Genome Project and Data Analysis Technology" Proceedings], 1991.

11. Hirosawa, Hoshida, Ishikawa, and Todani: MASCOT: 3 Jigen Dainamikku Puroguramingu Ni Motozuita Tampakushitsu No Araimento, Nihon Gakujitsu Kaigi "1992 Nen Johogaku" Shimpoijum Rombunshu [MASCOT: Protein Alignment Based on Three Dimensional Programming, Science Council of Japan "1992 Information Science" Symposium Proceedings], 1992.

12. Tanaka: Bunshi Seibutsugaku No Detabesu, Joshorigakkai Detabesu Shisutemu Kenkyukai [Molecular Biology Database, Information Processing Society of Japan Database System Forum], 84-23, pp 191-200 (July 1991).

13. Tanaka, H.: Protein Function Database as a Deductive and Object-Oriented Database, *Database and Expert Systems Applications*, Springer-Verlag, pp 481-486 (August 1991).

Research on User Interfaces in Massive Genome Analysis

92FE0879L Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 37-38

[Article by Masami Hagiya, Mathematical Analysis Laboratories, Kyoto University]

[Text]

Background and Goals

The research representatives have conducted research on developing environments for widow systems and user interfaces, especially graphic user interfaces for many years. At present we are conducting research on visual programming applying automatic programming methods. We are particularly interested in methods of visually expressing inference processes such as mathematical proofs. Our goal is to make use of this experience to conduct research on user interfaces for the purpose of massive genome analysis. More specifically, our present research goals are:

1. Visual user interface possibilities for genome analysis.
2. Realization of a user interface with capability for transfer of massive literature databases.

The first goal is an attempt to apply the above visual programming research to user interfaces for genome analysis. However, as mentioned above, this attempt has many unknowns. Therefore, as a more realistic research theme, we decided to work on the realization of a user interface with capability for transfer of massive reference databases.

Details of Study

For the first goal, we performed studies on visual interface possibilities for genome analysis. For the second goal, we studied realization of a relational database user interface on the standard universal text editor GNU Emacs.

Results

We found there are the following possibilities for the first goal.

1.1. Genome Database: Genetic maps and chromosome bands will be used to identify the positions of genes on the database.

1.2. Alignment Motif Search: Coloring of motifs, color coding of types of amino acids, and designations of intron positions will be performed. Moreover, a homology matrix can visually express homology.

1.3. Protein Structure Display: Visualization of primary structure, secondary structure, tertiary structure, and tertiary three dimensional structure is actively being studied in fields such as X-ray crystallography, NMR, molecular design, drug-receptor interactions and the like.

For the second goal, we are presently working on a user interface for SYBASE using the Emacs of GNU Emacs. Particularly, we have learned that it is important to create a routine interface that biologists can use.

Discussion

Concerning 1.1 and 1.2 above, we must admit that they are too primitive for visual interfaces. We must consider the following questions in the future: Are there no metaphors that use icons like those in office automation equipment? Further, are there no visual interfaces of a higher level than those that use icons for motifs?

Item 1.3 is the most important visual interface. However, we believe it is important not merely to make it visual, but to pursue the possibility of dialogue. For example, what about the possibility for a high level visual interface using artificial reality?

Concerning the second goal, we found that it will be extremely easy to create a routine interface that biologists can use by employing the customizing functions of a universal text editor that is generally considered standard by computer scientists. In the future we must ask: If this interface is realized, how will its quality compare to commercial items for use on personal computers?

References

1. Hagiya, Masami: Shikakuteki Puroguramingu To Jido Puroguramingu, Konpyuta Sofutouea [Visual Programming and Automatic Programming, Computer Software], Vol 8, No 2, 1991, pp 27-39.

2. Masami Hagiya: A Formal Framework for Visual Proving Based on Logic Programming, submitted to IEEE Workshop on Visual Languages '92.

Research on Methods of Expressing Mutation Spectra

92FE0879M Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 39-41

[Article by Yuzuru Fushimi, Hisashi Sato, Miho Suzuki, and Tadaaki Saito, Engineering Dept., Saitama University]

[Text]

Background and Goals

A biological species is not an existence characterized by a single sequence of DNA bases, but is an assembly of mutations. Each mutation is located at a single point on base sequence space, and a biological species corresponds to a localized distribution of these points on sequence space. Knowledge of the nature of this distribution is important when we estimate the characteristics of the species as a whole from small numbers of sequence data, and when we study the mechanism of molecular evolution.

A mutation spectrum is an attempt to establish and express links between the distribution of genotypes as base sequences and the distribution of phenotypes, which provides a general evaluation as a selection coefficient, within a certain biological population. The ultimate goal of this research is to create such a knowledge base.

As we refer to it here, what is a population? We would like to say that it is a sample group that reflects the population of a certain biological species as a whole, but random sampling from nature is impossible, and it is difficult to isolate the effects of different environments from the genetic effects. For example, many mutants of hemoglobin are known, and a database could be created, but the vast majority of these instances result from clinical screening, and the sample population would be biased toward abnormal mutations in the sense that these mutants are inferior to normal hemoglobin. Further, statistics on intentional modification resulting from protein engineering are difficult to relate to the general population since those intentions are biased.

On the other hand, mutant populations obtained from random mutations by random DNA synthesis and from the process of evolutionary molecular engineering are thought to reflect a spectrum (natural range) under a specific environment without bias. In this recently established field of evolutionary molecular engineering (see References 3 and 4 below), a mutation spectrum database can be created. We will develop a method to express those statistical features. What should be emphasized in statistical features is, for example, the extent of correlation between point mutations. We will compare this kind of unbiased data with features of biased data from nature.

Details of Study

To get a grasp on devising a method for analyzing the characteristics of a mutation spectrum in sequence space,

we must first attempt to express it visually. Therefore, this year we studied the geometric characteristics of base sequence space itself, selected from those various methods of expression the one that appealed most to the visual sense, and then studied methods to express mutation spectra in six dimensional base sequence space that is visualized in that manner.

Experimentally, it is necessary to develop a method for detecting differences in base sequences from a large quantity of clones. We studied temperature sweep gel electrophoresis (TSGE) as one such method.

Results

Geometric Expression of Base Sequence Space

The set of all character strings with a length v comprised of two letters (0, 1) is called 2 value v dimensional sequence space. We can introduce a distance concept to this with the Hamming distance d . The distance between two points is the number of different characters between two points; that is, the number of point mutations joining two strings of characters. It is well known in information science that if we join the points where $d = 1$ with line segments, we can express this sequence space in v dimensional volume (that is, $(v + 2)$ dimensional super line segments). On the other hand, we can view DNA/RNA as character strings comprised of four characters (A, T, G, C), so we can express DNA with a length of v bp as a single point in 4 value v dimensional sequence space (base sequence space). If we join the points where $d = 1$ with line segments in the same manner as in the above case with two values, it can be expressed by a super tetrahedron of $(v + 2)$ dimensions. We call this a connectivity map.

The connectivity map enables expression of the proximity relationship of mutations, and as v increases it becomes extremely complex and not suitable for visualization. The rules for constructing a connectivity map are fractal, and the construction procedure is drawn as a fractal graph. If a unit regular tetrahedron is projected onto a plane from a direction such that it is projected as a square, a map called a construction map can be obtained. There are $v!$ types of construction maps corresponding to the number of permutations when v coordinates are expanded. The construction map efficiently utilizes a two dimensional plane, which is the basis for visual sensation, the addressing is not very complex, and if the above permutations are chosen well, it can express mutation proximity relationships to a certain extent rather well (see Reference 5). In comparison, methods that display on a line or circle using conventional trees and gray code have quite poor efficiency for utilizing the visual sense.

Mutation Spectrum

The ultimate goal of this research is to create a knowledge base of mutation spectra, which link the genotype distribution and phenotype distribution in a certain population, but at present there are none available for use. First, for the purpose of simplicity, we dealt with phenotype in two forms, merely as existing or not existing. In this case, the

spectrum becomes the numerical distribution itself of the genotype (character strings) within a certain population (set of character strings).

Using the data of T. Schneider et al., in which they obtained 53 strong promoters from among 6,100 clones as a result of random mutagenesis of the T7 promoter $\phi 10$, we plotted the -14 to -9 regions on a six dimensional base sequence space construction map. We compared this with a plot of the -10 regions of 263 promoters of *E. coli*. We found it strikingly clear that the distribution has an amplitude of about 1 Hamming distance in the former and about 2 in the latter. This method of expression is more convenient for knowing the mutual relationships among loci than methods that express numerical distribution or existence ratios with respect to Hamming distance on a circle graph, or methods that express the ratio of consensus characters for each locus. For example, when we plot the results of C. Tuerk et al., who optimized the binding site sequence in T4-DNA polymerase binding RNA by RNA evolutionary molecular engineering, we come up with two localized regions. Moreover, we can clearly see that the mutually related loci can jump between both localized regions if simultaneous base substitution occurs. One of the two localized regions matches that discovered in nature. The jump between two localized regions is not so difficult, so there is a possibility that the latter will also be discovered in nature (see Reference 5).

Further, we are attempting to construct a system using SYBASE on the SUN workstation for the evolutionary molecular engineering mutant database, but we have not yet reached the publication stage.

Experimental Detection Method for Mutants

In DNA where a point mutation has occurred, the melting temperature T_m of the cooperative melting region that contains the mutation site will change slightly. If a GC clamp is applied by PCR next to this cooperative melting region, at temperatures near T_m the mutant DNA remains as a double helix, but other mutant DNA undergoes partial denaturation. This difference can be observed as a shift in mobility in gel electrophoresis. In the past this was observed on a gel with a temperature gradient (TGGE). We observed a similar phenomenon in a method that sweeps the temperature during electrophoresis (Temperature Sweep Gel Electrophoresis). In many cases the theoretical resolution in TSGE is slightly better than in TGGE. With this method, we could detect a single base substitution in 260 base pairs (see Reference 1). Further, as a method to perform rapid sequencing of large amounts of DNA, we developed a method for rapid enzymatic synthesis of primers that utilizes a library of DNA tetramer blocks (see Reference 2).

Discussion

We are still at the preliminary stages of the study, but this method gives an indication of the development of statistical analysis methods of correlations between mutations.

References

1. Yoshino, K., Nishigaki, K., and Husimi, Y., Temperature Sweep Gel Electrophoresis: A Simple Method To Detect Point Mutations, *Nucleic Acids Res.*, 19, 3153 (1991).
2. Kinoshita, Y., Nishigaki, K., and Husimi, Y., Enzymatic Synthesis of Sequencing Primers Based on a Library of Tetramers, *Chemistry Express*, 7, 149-152 (1992).
3. Fushimi, Y., Jikkenshitsu Ni Okeru Bunshi Shinka, *Kagaku* [Molecular Evolution in the Laboratory, Science], 61, 333-340 (1991).
4. Fushimi, Y., Shinka Bunshi Kogaku No Hajimari, Seibutsu Butsuri [The Beginnings of Evolutionary Molecular Engineering, Biophysics], 19, 22-25 (1992).
5. Fushimi, Y., Hairetsu Kukan To Totsuzen Heniitai Supekutoru No Hyogenho No Kento, I, Dai 2 Kai Kokai Wakushoppu "Hito Genomu Keikaku To Joho Kaiseki Gijutsu" Rombunshu [Study of Methods for Expressing Sequence Space and Mutation Spectra, 2nd Public Workshop "Human Genome Project and Data Analysis Technology"], Proceedings, 90-93 (1991).

Creation of a Database for Experimental Data on Human Genome Analysis and Its Applications

92FE0879N Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 42-46

[Article by Akisao Fujiyama, general director of Genetic Research, National Institute of Genetics]

[Text]

Background and Goals

We can say data processing performed by experimental researchers has two aspects. The first is data processing as a means to aid in the production of experimental data. The second is data processing to obtain a higher level of information from experimental data, for example, by comparing it with existing data. Regardless of whether a computer is used or not, research proceeds as a harmonious whole at the site of experimental research amidst mutual feedback from these two aspects. In the case of the latter, the use of computers has become general practice, but in the case of the former, the needs are various and we can assume that in most laboratories the processing is performed manually. However, we can foresee that when the object of analysis is huge, as in the case of genome analysis, not only do the types and quantities of research resources that must be handled increase astronomically, but also the experiments themselves generate huge amounts of data. It is impractical to deal with this kind of situation manually, and it becomes necessary to have computers (software) that have been developed for the purpose of supporting experiments and can be used easily at the laboratory level. In this research plan, we studied research support systems that utilize computers and the requisite conditions under which (molecular) biologists can use a workstation or a roughly comparable high level

personal computer at benchside for the purpose of supporting experiments and primary data processing. Our work was based on projections of the various experimental data and experimental resources that will be produced from genome mapping experiments that the reporters themselves are planning and carrying out.

Details of Study and Results

I. Construction of Genome Database With Macintosh

Because the system is for use at the laboratory level, the primary requisite condition is its ease of use. The greatest characteristic of the Macintosh computers installed with this year's allotment of research funds is that they are user friendly, so they are suitable for this purpose. At present we are linked to domestic and foreign networks via the National Institute of Genetics computer (ddbj.niguts), and we can use the GDB at Johns Hopkins University and the GBASE at Jackson Laboratories as genome-related databases.

First we decided to collect genome-related experimental data produced by individuals in separate labs, and make a database for genome-related data and a database for various types of data concerning cancer-related genes on the Macintosh so that they can be easily used for comparisons with

existing data. At present, the GDB noted above is open to the public as a human genome-related database, but even though we are linked by a network, throughout the day the responsiveness is slow, and it is practically useless. Under the present version, we can only search the GenBank ACC# for individual PROBE and PCR sequences, so the link with DNA sequence data is weak. Further, when we saw a sequence linked to a specific gene, the exon data, etc., were not included. The annotation is also weak compared with the DNA Data Bank and Protein Data Bank (Example 1). When data is needed we can access OMIM, but data transfer service from OMIM is not supported. In light of these problems and the limitations of using the Macintosh as hardware, we created the following basic specifications.

(1) The premise is use on an individual level. The search contents will be: data on gene loci and functions, data on specific regions on chromosomes, gene names and enzyme names, DNA base sequences and the like. We will make part of the data related to genome research, including GDB, easy to access. Therefore, we selected appropriate data entries to be recorded rather than all the GDB data.

(2) By using the OMIM notation rather than GDB annotation, the added value of the data will be increased.

| LOCUS MANAGER - GENERAL USER | | X |
|------------------------------|-----------|---|
| Go To | View | * |
| Call | Retrieve! | ? |
| Output! | | |

| | |
|---|----------------|
| Locus 1 of 1 | |
| Symbol: ALPL | Prev Symbol: |
| Name: alkaline phosphatase, liver/bone/kidney | |
| EC Num: 3.1.3.1 | |
| MIM Num: 171760 | |
| Location: 1p36.1-p34 | |
| Mode: A,L,R,S | Cloned: Yes |
| Status: C | Polymorphic: + |
| Het: .495 | PIC: .372 |
| Annotat: | |

| | |
|--------------------------|---------------------------------|
| Created: 01 Jan 86 00:00 | Fully Approved: 22 Aug 90 20:58 |
|--------------------------|---------------------------------|

Example 1. Output Screen From GDB

(3) We will fully supplement data related to DNA base sequences.

(4) We will make it possible to record and organize data on the individual and laboratory level, and make it possible for data required for individual research to be added at any time.

(5) By using database software for the Macintosh (4th DIMENSION), we will take advantage of the ease of use of the Macintosh. We are also considering making it relational.

Based on the above approach, we set to work. The present situation is as follows:

(1) After collecting and organizing data related to chromosome 2 from GDB and GenBank/DDBJ, we stored it in 4th DIMENSION as flat files. There were a total of 222.

(2) We took LOCUS, DEF, ACC#, KEY and COMMENT of sequences linked to each gene from GenBank and recorded them together with the GDB data.

• Items originally with ACC# in GDB:

• 42 Items newly searched and taken from GenBank: 83

(3) We recorded 155 items of data concerning cancer-related genes from GDB. Further, we recorded OMIM records concerning these genes (Example 2). We also plan to take human-related items from the DNA data bank and record them.

(4) We are in the process of creating a file from GDB compiled only from the records with MIM numbers (Example 3). This will be especially useful when MIM data is needed. With regard to this, as of this writing in February 1992, there are 12,223 records in GDB. Among these, records with MIM numbers total 1,959.

In the future, we plan to proceed with work on recording data and continue with studies of screen design, making data relational, and formats for recording image data.

| GDB-MIM | | | | | |
|---|---|----------------|-----------------|--------------|---------|
| Symbol | ALPL | LOCUS | 1p36.1p34 | | |
| Name | alkaline phosphatase, liver/bone/kidney | | | | |
| MIM Num | 171760 | Cloned: | Yes | EC Num | 3.1.3.1 |
| <p>PHOSPHATASE, LIVER ALKALINE [ALPL; ALKALINE PHOSPHATASE, LIVER/BONE/KIDNEY TYPE]</p> <p>Harris et al. (1974) found no genetic variants by electrophoretic means. However, the existence of at least one gene coding for the liver, bone, and kidney forms, independent of the other forms, is inescapable (Goldstein et al., 1980).</p> <p>Fedde and Whyte (1990) demonstrated that the alkaline phosphatase of skin fibroblasts is ALPL, the tissue-nonspecific type, and that it is active toward millimolar concentrations of the putative natural substrates phosphoethanolamine (PEA) and pyridoxal-5-prime-phosphate (PLP). Both activities were deficient in hypophosphatasia fibroblasts. They presented evidence that normal fibroblast ALP is linked to the outside of the plasma membrane; thus, the enzyme acts physiologically as a lipid-anchored PEA and PLP ectophosphatase.</p> | | | | | |
| PIC: | .372 | HET: | .495 | Polymorphic: | + |
| Status: | C | | Mode: | A,L,R,S | |
| GDB Annotation | | | | | |
| Created | C 1 Jan 86 00:00 | Fully approved | 22 Aug 90 20:58 | | |

Example 2. GDB-MIM

MIM Num 171760

Annotation

PHOSPHATASE, LIVER ALKALINE [ALPL; ALKALINE PHOSPHATASE, LIVER/BONE/KIDNEY TYPE]

Harris et al. (1974) found no genetic variants by electrophoretic means. However, the existence of at least one gene coding for the liver, bone, and kidney forms, independent of the other forms, is inescapable (Goldstein et al., 1980).

Fedde and Whyte (1990) demonstrated that the alkaline phosphatase of skin fibroblasts is ALPL, the tissue-nonspecific type, and that it is active toward millimolar concentrations of the putative natural substrates phosphoethanolamine (PEA) and pyridoxal-5-prime-phosphate (PLP). Both activities were deficient in hypophosphatasia fibroblasts. They presented evidence that normal fibroblast ALP is linked to the outside of the plasma membrane; thus, the enzyme acts physiologically as a lipid-anchored PEA and PLP ectophosphatase.

Swallow et al. (1985, 1986) used a monoclonal antibody to distinguish

human from rodent forms of the 'liver/bone/kidney' isozyme of alkaline phosphatase, the isozyme deficient in hypophosphatasia (146300, 241500). In human-rodent somatic cell hybrids, segregants indicated that the human ALPL locus is on chromosome 1. Mouse Akp-2, which may be homologous, is on chromosome 4 between Pgm-2 and Pgd. Thus, ALPL might be located on 1p between PGM1 and PGD. (It is not completely certain whether the mouse homolog is Akp-2 or Akp-1; the latter is on mouse chromosome 1 (Nadeau, 1988).) Weiss et al. (1986) cloned cDNA for ALPL. ALPL showed 60% homology to placental alkaline phosphatase (171800) over the protein coding part. Weiss et al. (1988) and Matsuura et al. (1990) found that the ALPL gene exists in single copy in the haploid genome and is composed of 12 exons distributed over more than 50 kb. As compared with the gene isolated using a bone-type ALPL cDNA (Weiss et al., 1988), liver-type ALPL mRNA was found to have another leader exon about 3.4 kb upstream

from exon 2 and alternative splicing in the first exon was indicated (Matsuura et al., 1990). Using a DNA polymorphism of the ALPL gene, Ardinger et al. (1987) found linkage to Rh (theta = 0.08; lod = 5.50). Multipoint analysis indicated the following order: Rh--3--ALPL--12--GLUT--23--PGM1, with the interlocus intervals as percent recombination in males (the female rate about 2.8 times the male rate). Weiss et al. (1987) assigned the ALPL locus to 1p36.1-p34 by demonstrating linkage to Rh (maximum lod score = 5.50, theta = 0.10). In the full report, Smith et al. (1988) established the location in 1p36.1-p34 by a combination of Southern blot analysis of hybrid cell DNA, in situ hybridization, and genetic linkage analysis. Linkage to Rh was indicated by a maximum lod score of 15.66 at a recombination fraction of 0.10 and to fucosidase A (230000) by a maximum lod score of 8.24 at a recombination fraction of 0.02. Both Greenberg et al. (1990) and Kishi et al. (1991) did prenatal diagnosis using

linked DNA markers. In the family they studied, Kishi et al. (1991) pointed out that serum alkaline phosphatase activities in the mother, a heterozygous carrier, fell within the normal range, while those in the father and paternal grandmother were below normal. Urinary

II. Use of Computers for Experiment Support

Concerning a support system for mapping experiments, preparation on the experimental end is running behind, and we are still at the planning stage. We are thinking about individual data storage and making private databases, clone sequences (creation of contiguity and clone islands) and the like. However, with the exception of one research group that has started large scale mapping and sequencing, for the time being we think that small scale, easy to use systems like the Macintosh are useful. Further, we are studying problems such as what kind of data we will record from experiments and what will be required in the future.

Discussion

Software for large scale mapping is being developed independently at the laboratory level now in the United States and other countries. Further, even though computers are installed, there are many experimental methods built on the premise of analysis by the eyes of the researcher, and it would be impractical to computerize them, so it is necessary to propose experimental methods built initially on the premise of performing the analysis by computer.

Concerning the human genome database, for the sake of reference there are 53 researchers (or groups) in Japan that have accounts on GDB, but among these there are fewer than 10 that use it actively. Possible explanations are that the absolute number of researchers is small, they are not known, they do not feel a particular need for a DNA sequence database right now in terms of their present research, they feel that GDB cannot be used in performing searches for unknown data because of the features of the recorded data, and they believe that the data is mainly data on chromosome loci. We believe that the time is near for an integrated database based on genetic maps.

Including the above topics, if genome analysis is to proceed smoothly and the data thus obtained is to be fully utilized, the cooperation of information science specialists will be indispensable.

Structure of Knowledge Base for Transmembrane Proteins

92FE08790 Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 47-50

[Article by Shigeki Mitaku and Makiko Suwa, Engineering Dept., Tokyo University of Agriculture and Technology]

[Text]

Background and Goals

Learning the three dimensional structures and functions of proteins by the amino acid sequences of structural gene segments is one of the major goals in the analysis of genome data. To accomplish this, it will be necessary to construct amino acid sequence analysis systems using a variety of knowledge from physical chemistry and biology. This research focuses on transmembrane proteins, which comprise an extremely important category among the proteins. Our goal is to perform the following series of

studies to construct a system for analyzing amino acid sequences based on a knowledge of physical science.

(1) We will establish a method for categorizing amino acid sequences into three categories (A: transmembrane proteins that are almost entirely within the membrane and have a high helix content, B: transmembrane proteins that are almost entirely in the water phase but have portions that span the membrane, C: other proteins, including water soluble proteins).

(2) Utilize graphics, energy calculations, etc., to clarify the relationship between the primary structure and three dimensional structure of transmembrane proteins in category A.

(3) In many cases the transmembrane proteins in category A bind and function within the surface of the membrane, and it is thought that they have common structural conformations and functional mechanisms. By studying the relationship between functional activity of mutants and energy calculations, we can estimate the type of function.

The subjects of this study are transmembrane proteins, which is only a partial set of proteins (a numerical percentage of all proteins), but the goal of this study is to estimate structure and function by obtaining high quality knowledge data about these extremely important transmembrane proteins.

Approach and Details of Study

The details of the study can be divided roughly into three categories. More specifically:

(1) We will distinguish transmembrane proteins by analyzing the hydrophobicity of their amino acid sequences. In other words, transmembrane proteins are composed of bundles of several membrane-spanning helices bound together. Because the membrane-spanning helices are buried within the non-polar lipid membrane, most of these helices are extremely hydrophobic. Or conversely, because they are hydrophobic, they can enter the membrane and form a stable structure. With this knowledge, for the first module we adopted a method of distinguishing types of proteins by determining the total amino acid hydrophobicity from the hydrophobic values of amino acid sequences and the periodicity of their long periods. This is already operating on the personal computer level, and we have begun to transfer it to a UNIX workstation to link it with other modules.

(2) Predictions of protein secondary and tertiary structure have been partially conducted using empirical methods, but because very few three dimensional structures of transmembrane proteins are known, there has been no clear method for predicting structure. We performed a variety of experiments (including computer experiments) based mainly on denaturation, and it became clear to us that polar interactions are the true essence of transmembrane protein tertiary structure stability. Therefore, we have begun to predict structure with energy calculations that incorporate polar interactions. At this time when we performed calculations with resolution at the atomic level, we found that minute local minima of energy interfered, and conversely, we could not see the main protein configuration (helix location, etc.). Therefore, we thought that it was important to calculate energy profiles at a resolution of about 1 nanometer, for example, and we attempted this with an energy calculation method that uses a probe helix.

This is the second module. Many probe helices are possible, and by changing the probe helix it is possible to change the spatial resolution of the energy profile. After the helix position is determined, when we consider even more detailed structure, we use the dual method of increasing the resolution by working with the probe helix, and calculating and minimizing the total interaction energy to actually position all the membrane-spanning helices.

Further, for the loops that are believed to be exposed to water, we adopted the ideas of Saito et al. (Waseda University) that hydrophobic interactions are important in the tertiary structural conformation of water soluble proteins, and attempted to predict structure.

(3) After we learn the three dimensional structure, then it is important to consider what kind of data processing is possible in terms of physical chemistry. As part of this flow, we are considering a system to study the relationships between functional activity of mutants and structural energy calculations. Basically, we will use a commercially available application, but for this problem it is important to begin by finding parameters that show a good correlation with activity such as using the total energy or the interaction energy. At present this is under study using bacteriorhodopsin as the specimen. We are considering adding a portion of software to a commercially available application so that it can solve this kind of problem easily. This is the third module.

To perform these calculations we used an IRIS 4D/35 graphics workstation from Silicon Graphics (hardware) and CHARM and QUANTA from Polygen [phonetic] (software).

Results

This year we proceeded with research mainly on the second module. We decided to solve two problems in

predicting the tertiary structure of transmembrane proteins. First we decided to study the characteristics of interactions around each membrane-spanning helix, and second we decided to solve a kind of jigsaw puzzle by neatly combining those characteristics. We used the probe helix method as the method to study the helix characteristics while slowly reducing the resolution. In this method we generated the helices to be studied and the additional probe helix on a graphics workstation and calculated the interaction energy for the system. When we systematically changed the arrangement between the probe helix and the helices we were studying, we could learn how they interact in a number of helix directions.

For the seven membrane-spanning helices of halorhodopsin, we made calculations using polyserine as a probe helix, and the results are shown in Figure 1. The irregular line on the interior of the wheel shows the interaction energy values, and the black portions show portions with striking polar interactions. It is clear that for each helix there are portions facing the adjacent helices and portions with strong polar interactions in the direction of the protein's interior. This confirms the implication from experimental results that the helix portions penetrating the membrane are mutually bound by polar interactions.

However, with this alone the data is insufficient for predicting structure (spatial resolution is too poor). To overcome this problem, it will be necessary to work on the probe helix. The next probe we tried was a helix composed of serine and alanine, and we came up with a copolymer helix, for example, with serine on the upper portion and alanine on the lower portion. When we made the calculations, we found that only the polar interaction characteristics of the upper portion were evaluated. When we tried this with halorhodopsin, the profiles with the upper and lower portions were clearly different. For example, helices

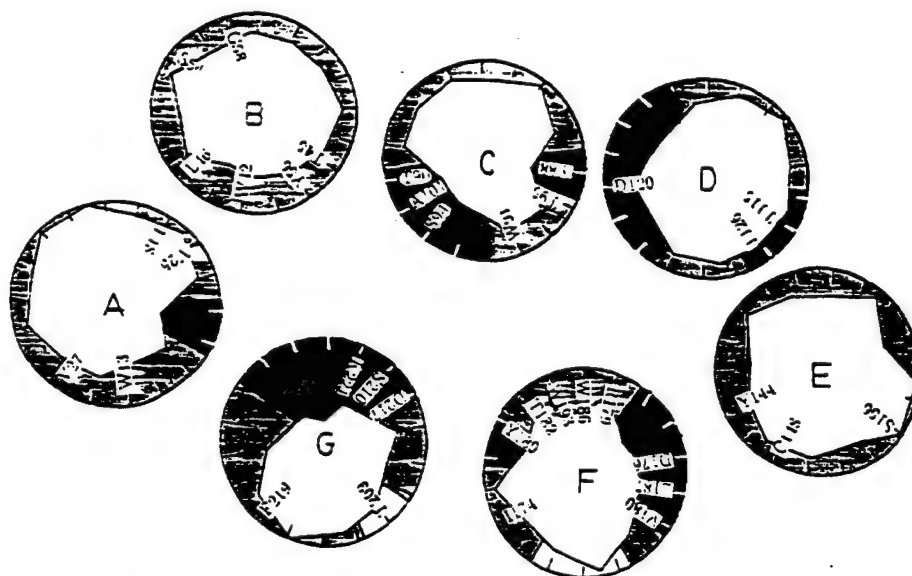


Figure 1. Evaluation by Probe Helix Method of Halorhodopsin Membrane-Spanning Helices

A, B, C, and G showed clear polar interactions with the lower portion of the helix, and conversely, helices D, E, and F showed strong polar interactions with the upper portion. This clearly shows that the energy is low and therefore advantageous when helices A, B, C, and G, and helices D, E, and F form their own little groups. If that is the case, then we can reduce the combinations of helix locations from 5,040 (7!) to 144 (3! x 4!). The results up till now are still inadequate, but they suggest that if we increase the resolution somewhat, we can determine the unique helix arrangement.

On the other hand, the jigsaw puzzle problem requires using approaches from information science. We are considering using a neural network and are studying joint research with Dr. Akiyama from the Electrotechnical Laboratory.

Conclusion

We have attempted to construct a software system consisting of three modules with the goal of obtaining high level information from the standpoint of physical chemistry concerning the three dimensional structure and function of proteins from their amino acid sequences. We have basically completed the first module and are in the process of transferring it to a UNIX workstation. For the second module we divided the problem into an evaluation of the characteristics of helices that uses the probe helix method and a three dimensional jigsaw puzzle problem. We established the methodology of the first part and proceeded to make specific calculations using a protein with seven helices as a specimen. We will improve the program to make the system easier to use. For the third module we have just finished the most basic concepts of the system, and research results are still insufficient for considering software for specifically dealing with individual transmembrane proteins. We will continue to study this next year.

References

1. Makiko Suwa, Shigeki Mitaku, Kazuko Shimazaki, and Tatsuji Chuman, "Characterization of Transmembrane Helices by a Probe Helix Method of Molecular Energy Calculation," JPN. J. APPL. PHYS., 31 (1992) in press.

Structure and Integration of Japanese Language Human Genetic Map (JHGM) Database

92FE0879P Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 51-54

[Article by Shinsei Minoshima, Dept. of Medicine, Keio University]

[Text]

Background and Goals

Human genome mapping data is compiled each year at the Human Gene Mapping Workshop (HGM) and recorded in the GDB (Genome Database). The GDB is maintained and managed by Johns Hopkins University in the United States. The present amount of data in the GDB, including results compiled at the HGM Workshop (HGM 11) held in

London last August, has reached 12,500 items that includes DNA fragments and 3,000 gene entries. The GDB is open to the public so that researchers all over the world can use it via a computer network, and we can also access it from Japan via a telephone line and Internet. The GDB provides a common worldwide format for compiling various kinds of mapping data, and it actually integrates and manages the data as a whole, so it is very important. However, there are some problems because GDB is remote, and especially when we use telephone lines from Japan, we cannot overlook the delay in access speed and telephone costs. Moreover, the fact that it cannot handle data in Japanese limits its users. It is highly significant that we create a genetic mapping database separate from GDB that can also handle descriptions in Japanese and can be used easily in the laboratory. The JHGM/PC (Japanese Human Gene Mapping Library on PC) database that we created in the past runs on NEC personal computers and Toshiba laptop computers that are widely used in Japan. As a database management system (DBMS) it uses LET'S IRIS (by Personal Media) on MS-DOS. It describes gene traits in Japanese and inputs personal map data and reference data. Our main goals for this year are the following three items.

1. Study and improve JHGM/PC data items.
2. Revise and input JHGM/PC data.
3. Transfer JHGM/PC to a workstation.

Details of Study

1. Study and Improve JHGM/PC Data Items

(1) The original version of LET'S IRIS, which the JHGM/PC uses as a DBMS, was 2.0, but it has now been updated to version 4.0. We will convert the data and customize it so it can be used on the new functions such as the screen displays.

(2) We will add genetic traits in English as a new item in the database.

2. Revise and Input JHGM/PC Data

The total data numbers for JHGM/PC at the end of last year were 6,316 map entries and 10,279 reference entries. We will revise and input new databases on the results of HGM 11 that was held in London in August last year.

3. Transfer JHGM/PC to a Workstation

JHGM/PC is easy to use, but because the quantity of data has increased, the search speed has dropped. Moreover, just as in the case of GDB when used by remote, there is room for improvement in the user interface. With these things in mind, we are trying to develop a new gene map database library system called JHGM/X (Japanese Human Gene Mapping Library on X-Window).

- a. Operates on general X-Window.
- b. Runs rapidly with the whole database on memory.
- c. Handles Japanese data according to EUC codes.
- d. Has a smooth interface using a mouse as a pointing device.

e. Displays chromosome bands graphically as patterns, and map displays are much more visual.

Results

1. (1) Conversion of JHGM/PC data from LET'S IRIS 2.0 to LET'S IRIS 4.0 went smoothly. Revising the index file took some time, but there were no problems. Concerning the main features on the new version, because the screen display layout card can be customized, we positioned both map and reference items on the screen so they can be used effectively.

(2) We changed the LET'S IRIS card format so that we could add the item for genetic traits in English. This also went smoothly.

2. This year's input data includes 250 new map cards and 4,851 new reference cards, and when we add this to last year's cards, the total is 6,566 map cards and 15,130 reference cards. In addition when we include the genetic traits in English mentioned above, more detailed data on genes already mapped and revisions in reports that we added to and revised in last year's cards, the amount of data input was actually greater than what was mentioned above.

3. The JHGM/X screen we created is shown in Figure 1 [not reproduced]. We developed our own JHGM/X data processing portion with C language. We described the control portion for the screen displayed on X-Window with the Athena Widget Set developed at MIT. The data uses two flat data files (genetic map data and reference data) with the EUC codes converted from the JHGM/PC. The genetic map data file has fields for gene symbol, traits, gene locus, etc., and the reference data file has fields for gene symbol, author, title, journal name, etc. We linked these two data files using the gene symbol as the common key. When the system starts up, a single data structure consisting of multiple structures is compiled and expanded into virtual memory, so very high speed data searches are possible. Below, the functions of the JHGM/X are described.

a. Japanese Language Function

Just as in JHGM/PC, data in Japanese is used for the gene traits and comments. For these two items, Japanese can be used as the keyword in the searches described below.

b. Smooth User Interface

Selection and Operation by Mouse

Operation of the JHGM/X basically can be performed by giving it the minimum amount of data from the keyboard, such as a search keyword, and selecting a function button for the desired output form with the mouse.

Keyword Window

This window can be used as a word pool for storing keywords that the user uses with high frequency. Its contents can be stored and reused as a file in the directory for each user. If the keyword is input from this window by using the cut and paste function with the mouse, input from the keyboard can be reduced ever further.

c. Search Functions

Two main types of searches can be performed on the JHGM/X.

Genetic Map Search Function

Genes or references can be searched for from the field consisting of map data. For simultaneous designation of multiple items, an AND operator can be used. In the gene symbol field, wild card symbols (?, *) can be used and in the gene trait field, multiple keywords and AND searches (connected by &&) can be used. The order in which search results are displayed can be selected by order of region or by alphabetical order of the gene symbols. It is also possible to limit the search object to either genes or DNA fragments.

Reference Search Function

A reference search can be performed from the field consisting of reference data. The main functions are in accordance with the genetic map search functions described above, and the order in which the search results are displayed can be either alphabetical by author or by year of publication.

d. Graphic Display of Chromosomes

A separate window (map window) is opened to display chromosome ideograms. The map window not only displays chromosome ideograms, but also operates as a graphic front end processor for searches. It has the following three functions.

Visual Display of Each Chromosome Ideogram

The displayed chromosome ideograms are drawn at a size that is relative to the actual chromosome. Further, because the centromeres of all chromosomes fall on the same straight line, it is possible to line up multiple chromosomes and compared their relationships easily.

Visual Display of Gene Regions

When a gene symbol is designated, the database is searched, the locus of the gene determined, and it is displayed with a bar and gene symbol at a corresponding position on the ideogram. Up to four genes can be displayed simultaneously in different colors.

Gene or Reference Search by Regional Designation From the Ideogram

It is possible to designate a region on the ideogram and search for genes that lie in that region and their references. There are two types of search methods depending on the size of the designated band.

Discussion

1. By using a new version of LET'S IRIS for the JHGM/PC database management system, the ease of use increased, but because of the increasing amount of data, the search speed has dropped. There are many labs that cannot use the JHGM/X, and we believe that there is still sufficient value in using JHGM/PC. Therefore, we think that a good method will be to create subsets such as separating by chromosome or separating gene data from DNA fragment

data, and place these files together with a comprehensive file of all the data on a hard disk so that they can be used independently as needed.

2. The JHGM/X was created to use both Japanese and English, and if the database is prepared with appropriate characters and fonts, it can be easily used in other languages as well. In the future we would like to develop JHGM/X as a polylingual application for using GDB data.

References

1. Minoshima, S., Dohi, H., Ishizuka, M., and Shimizu, N., X-Uindo No Ue De Ugoku Hito Idenshi Mappu Detabesu No Kochiku, Johoshori Gakkai Kenkyu Hokoku [Construction of a Human Gene Map Database Running on X-Window, Information Processing Society of Japan Research Report], 91 (74): 1-6.
2. S. Minoshima, H. Dohi, M. Ishizuka, and N. Shimizu, The human gene map database in Japanese: Development of a new software for workstation. JPN. J. HUM. GENET. In press.
3. Dohi, H., Minoshima, S., Shimizu, N., and Ishizuka, M., Nihongo Hito Idenshi Mappu Raiburari No Sekkei, Johogaku Shimpoujumu Rombunshu [Design of a Human Genome Map Library in Japanese, Information Science Symposium Proceedings], 1992, pp 15-23.
4. H. Dohi, S. Minoshima, N. Shimizu, and M. Ishizuka, JHGM/X: Japanese Human Gene Mapping Library on X-Window. In preparation.

Research on Applications of Super Parallel Computing on Human Genome Analysis

92FE0879Q Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 55-58

[Article by Akinori Yonezawa, Science Dept., University of Tokyo]

[Text]

Background of Research

With advances in computer development technology in recent years, the super parallel computer has become a reality. This super parallel computer promises a leap to a higher level of performance over existing super computers in both the amount and speed of calculations because it has a large number of processors that process data simultaneously. In scientific and technical calculations, the use of the super parallel computer is expected to open the path to a new field of research in addition to the past utilization of computers mainly for number crunching.

The application of the parallel super computer to groups of genetic data such as nucleic acid sequences that continue to accumulate with progress in the Human Genome Project will be a desirable new area of research, judging from the vast quantity and homogeneity of the data.

However, not only in the case of genetic data but in other cases as well, a knowledge and understanding of computers, including hardware and architecture, is required, in addition to an understanding of the nature of the data to be

processed, to create software that will enable efficient implementation. There is a concern that the development of software on a parallel super computer, whose workings are more complex than previous computers, will be quite a burden on biological researchers, who are not computer experts. This research designed and installed a programming language to alleviate obstacles in the development of this kind of software that will accompany the introduction of super parallel computers to genetic data processing.

Details of Study

Super computing, which found practical application one step ahead of super parallel computing increases calculation speed by pipeline processing of matrix and vector operations in the program. On the other hand, the super parallel computer that we study uses a method in which the processing is divided up and allotted to multiple processors, and the processors proceed with the calculations in a coordinated manner. It is predicted that this method of calculation will be more suited to genetic data that includes not only numerical values but also large amounts of letter and symbol data. There are generally two methods for describing the parallelism of problems in programs under this kind of calculating environment. In one method the programmer himself actively describes parallelism included in the problem as a program, and in the other the computer automatically extracts parallelism from the problem that is described and performs parallel processing. We proposed the parallel object-oriented programming language ABCL for the former and the parallel conditional logic programming language PARCS for the latter, and we conducted research on their design and installation.

In the case of the former, to create software using an existing procedural programming language, we performed expansions of functions such as library coefficients for describing parallelism. Generally speaking, however, if the method of dividing the program is not appropriate, the efficiency is not greatly improved over sequential processing. Further, because the units running in parallel must interact during the calculations, exclusion control and synchronism of data structures for communication can become intricate quite easily and cause errors in the operation of the software. Errors generated by these kinds of low-level computer functions are extremely difficult to discover by debugging, so it is quite difficult to create software that operates well, and it requires a large amount of time. A parallel object-oriented programming language absorbs these kinds of difficulties in parallel processing descriptions at the level of the programming language and was designed for improving the productivity and reclamation of software. With this programming language, because the program is expressed as a set of objects that are the main subjects of the parallel operation, it is possible to reflect the natural parallelism of the problem at the level of program description. Furthermore, mechanisms such as synchronism and exclusion control are manipulated within the programming language processing system, and it is not necessary for the program to recognize them, so the health of the software is increased. We have proposed ABCL as a

language equipped with these kinds of features. This year we conducted research on the development and implementation of a processing system on an actual (super) parallel computer.

For installation we used the EM-4 developed by the Electrotechnical Laboratory and the AP-1000 developed by Fujitsu Laboratories. On the EM-4 we used an architecture specializing in the data flow calculation model, and on the AP-1000 we used an architecture that is generally used in workstations and the like. Under an environment that added rapid communication functions between processors, we conducted research on installing a programming language processing system that would enable efficient parallel running of the program.

Further, as an application of the parallel object-oriented language, we designed an object-oriented algorithm for multiple problems with similar calculations that will enable expansion to problems such as the study of molecular dynamics, and installed ABCL/1. In molecular biology, high-speed simulations of the behavior of giant molecules such as proteins inside the body will be useful in studies of high level structure and chemical characteristics. In the past, however, calculating power has been inadequate even when computers with the highest level of performance, such as super computers, were used. Therefore, we will continue to study simulation methods using the ABCL processing system on the above parallel super computers using this algorithm.

We proposed a model for a program language called a parallel constraint logic language as a method for automatically extracting the parallelism of a program, and based on that proposal we designed a prototype of the PARCS programming language processing system. In a parallel constraint logic programming language, the programmer expresses problems to be solved as conditions between the various structural elements of the problem.

Therefore, it is possible to describe the program in a form similar to the style of the problem, and it is unnecessary to consider control structures such as IF or GOTO, substitution of variable values, manipulation of the pointer and the like as we must do when we use a procedural programming language such as FORTRAN or C. As a result, it is possible to create an errorless program relatively easily in a short period of time.

In the PARCS processing system we proposed, we enhanced the language specifications so that it can handle constraints in finite realms of symbols and integers, in addition to the Herbrand realm that is handled by conventional logic programming language. As a result the range of problems that can be expressed was increased, and we can expect even more practical utilization.

Further, PARCS automatically extracts parallelism (AND parallels, OR parallels) contained in the program and runs them in parallel. In this kind of separate processing, the priorities are divided up and in the process of searching for the solution, the programming language system automatically trims branches efficiently. Generally speaking, program operating efficiency becomes a trade off with the

descriptive power of the programming language, but with PARCS, because of this powerful branch trimming function, even in calculating environments with limited calculation power, such as with parallel computers having a relatively small number of processors, it is possible to install a programming language processing system in which the program operating efficiency drops very little despite its high descriptive power. In addition, language specifications are created while keeping in mind the operation of the super parallel computer, and a large increase in processing speed can be expected when using super parallel computing.

Further, thanks to this function, when handling problems that generate an explosive amount of parallelism, such as in combination problems, it is possible to hold down the parallelism by removing from the solution candidates whose combinations will not yield a valid result. By having the processing system learn the priority parameters, we can expect increased efficiency when repeating calculations on the same problem. (Because these functions make it possible to proceed with calculations in accordance with the available calculating power as the size of the problem to be handled increases, these are valid under the calculating environment of the super parallel computer.)

We installed the PARCS prototype system to make pseudo-parallel runs on a variety of workstations, and we confirmed that the features of our language model can be realized. Then we conducted research on installation in a super parallel and decentralized computing environment on the Intel parallel computer IPSC/2. The IPSC/2 is a parallel computer in which processors without shared memory are linked in a hypercube state. For our installation we aimed at a method with high portability in a parallel/decentralized calculating environment that is not dependent on a specific architecture. By creating this processing system, we confirmed the behavior of the processing system in an actual parallel computing environment, studied the relationship between an increased number of processors and higher processing speed, performed verifications such as the extent of software productivity and ease of debugging, and obtained results that supported our hypotheses. The language specifications of this processing system were virtually a continuation of the specifications of the pseudo-parallel edition of PARCS, and research is still progressing on the language specifications and expansion of functions under a (super) parallel programming environment.

Today, much research on the application of constraint programming is underway in fields related to artificial intelligence such as in processing of natural language, but gradually it is beginning to deal with more and more subjects. Actually, in addition to our group, research has been announced in *Science* that uses constraint programming to describe a system for predicting RNA three dimensional structure, and we can expect an increase in the future of the use of constraint programming for data processing in biology, and especially in genetics. In our lab as well, we aim to develop a high speed application for calculating nucleic acid structure by using PARCS for data

processing of nucleic acid sequences and the calculating power of the super parallel computer.

Discussion

Concerning the relationship between PARCS and ABCL, we think that high speed operation using the ABCL super parallel processing system is good for problems in which the details and parallelism of the algorithm are clear such as in detecting specific patterns from character string data for extracting protein motifs, and that when dealing with amorphous problems in which the details of the algorithm are still not known, it is good to use PARCS. We believe that because programming is easier than with procedural programming languages, using super parallel computing will increase software productivity in the field of biology, and that our research will aid in the advancement of biological research as a whole.

References

Concerning ABCL

1. Yasugi, M., Matsuoka, S., and Yonezawa, A., ABCL/1 on EM-4: A New Software/Hardware Architecture for Object-Oriented Concurrent Computing on an Extended Dataflow Supercomputer. To appear in Proc. of 6th ACM International Conference on Supercomputing, Washington, D.C. (1992).
2. Yasugi, M., Yonezawa, A., N-Tai Mondai No Heiretsu Obujekuto Shiko Arugorizumu, Sofutouea Kagakukai Dai 8 Kai Taikai Rombunsho [Object-Oriented Algorithm of N-Body Problems, Software Science Society 8th Annual Conference Proceedings] (1991).
3. Taura, K., Mauruchikonpyuta Jo No Heiretsu Obujekuto Shiko Kengo No Kokoritsu Na Jisso Ni Kansuru Kenkyu. Sotsugyo Rombun [Research on High Efficiency Installation of Parallel Object-Oriented Language on Multicomputers, Dissertation] (1992).

Concerning PARCS

4. Kobayashi, N., Matsuoka, S., and Yonezawa, A., Control in Parallel Constraint Logic Programming, Proc. of the Logic Programming Conference, Lecture Notes in Artificial Intelligence, Springer-Verlag (1992).
5. Nagazuka, M., Seiyaku Ronrigata Gengo No Heiretsu Keisanki Jo E No Jisso, Sotsugyo Rombun [Installation of Constraint Logic Language on Parallel Computers, Dissertation] (1992).

Public Subscription Research Groups: High Speed General Character String Searches by Intelligent Algorithm and Parallel Processing

92FE0879R Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 59-60

[Article by Hiroshi Imai, Science Dept., University of Tokyo]

[Text]

Background and Goals

In the massive knowledge data processing accompanying genome analysis, intelligent, high-speed processing of large quantities of data such as DNA and protein sequences is required. A high-speed algorithm for character string processing is needed to process sequence data as character strings and determine the biological meanings contained within them. Because this algorithm must not merely process symbols but must process signals that satisfy a variety of characteristics for determining biological meaning, it must include elements of knowledge processing.

In this research we considered the problems of processing character strings with added elements of knowledge processing within the context of basic character string processing problems, and we attempted to design and develop a high-speed algorithm to solve those problems. We also considered the role parallel processing might play in increasing speed even more.

Details of Study

A high-speed algorithm for character string processing is needed to process sequence data such as DNA base sequences and determine the biological meaning contained within them. In this year's research we addressed the shortest shared super character string problem and the longest shared partial sequence problem, which are representative problems in character string processing. We did not treat these problems merely as problems of computer science, but we conducted research to make use of the special characteristics arising from the use of DNA as an object of study.

Here the shortest shared super character string problem is the problem of determining the length of the shortest among the character strings (called super character strings) that are contained as partial strings within other character strings when a large number of character strings are given. This problem corresponds to determining an original base sequence by sampling and reading partial strings of an appropriate length from an extremely long string (for example, a string of 500 characters) and joining them together to determine the original base sequence when actually reading base sequences from DNA.

The longest shared partial sequence problem is a problem of determining the length of the longest shared partial sequence in each character string when k character strings are given.

Results

As noted above, in this year's research we addressed the shortest shared super character string problem and the longest shared partial sequence problem, which are representative problems in character string processing, and we conducted research to make use of the special characteristics arising from the use of DNA as an object of study.

We realized that the shortest shared super character string problem is a difficult NP-complete problem, and in this year's research we studied it from the viewpoint of how much easier the problem will become when part of the

biological conditions are added. Up to this point we have learned that we can only find the super character string that is formed last, and the problem does not get much easier. Further, we knew that although the problem is difficult when viewed from the standpoint of computer science, from the standpoint of an approximation algorithm, this problem is solved on a daily basis as a practical problem. If we ask whether it is necessary to determine which one is really the shortest among the shared super character strings, we can say that it is sufficiently meaningful to solve the problem by approximation. In that case, the problem for future study will be what kind of meaning shall we give to the method of evaluating the approximation solution method in computer science from the standpoint of genetic data processing.

In past research on the longest shared partial sequence problem, an effective algorithm is given for the case of $k = 2$, but when $k \geq 3$ only algorithms that require a lot of computing time are known. Actually when k is considered an input parameter of the problem rather than a constant, the problem becomes an NP-complete problem.

For this problem, we first focused on the fact that there are only four types of characters (A, T, G, C) in base sequences, and we designed an algorithm to solve the problem more efficiently in cases in which $k \geq 3$. This algorithm basically is from dynamic programming, but it rapidly determines the most appropriate solution without constructing all the tables used in dynamic programming. The degree of parallelism is roughly the same as with previous methods. This algorithm is superior not only in terms of speed, but in the size of its working area. We aimed for solving the problem within a sufficiently practical amount of time on a workstation even when k is large. Next year we will install this algorithm and confirm its validity.

Further, with respect to the longest shared partial sequence problem, we are studying ways to increase the speed even more by using the techniques of artificial intelligence search algorithms such as the A* algorithm in the search algorithm we develop.

Future Topics

This year's research consisted of a study of algorithm pre-design (shortest shared super character string problem) and algorithm design (longest shared partial sequence problem). In the future we must install the algorithm and perform computer experiments using it. Only the most basic of the characteristics specific to genetic data processing that we have studied up till now are in use. We plan to study them in greater detail and consider new characteristics in terms of algorithms from the standpoint of computer science. Ultimately, we plan to link our work to the development of a high-speed algorithm program that will help in genetic data processing.

References

1. Hakata, H. and Imai, H., Mojisu Ga Sukunai Baai No Fukusu No Mojiretsukan No Saicho Kyotsu Bubunretsui Mondai, Johoshori Gakkai Arugorizumu Kenkyukai [Problem of Longest Shared Partial String Among Multiple

Character Strings When Number of Characters Is Small, Information Processing Society of Japan Algorithm Forum], 1992 (in press).

2. Hakata, H. and Imai, H., Saitan Kyotsu Chobubunretsui Ni Kansuru Kosatsu, Denshi Joho Tsushin Gakkai Konbyuteshon Kenkyukai [Thoughts on Shortest Common Shared Super Partial Strings, Institute of Electronics, Information and Communications Engineers Computation Forum], 1992 (in press).

Genome Description by Formal Grammar

92FE0879S Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 61-62

[Article by Yoshiyuki Kotani and Nobuo Takiguchi, Engineering Dept., Tokyo University of Agriculture and Technology]

[Text]

Background and Goals

Formal grammar logic is a method of handling strings of symbols numerically. It is presently used in a wide range of areas such as in the processing of natural language including comprehension of natural language, machine translation, and word processor document processing, in artificial languages including program languages, in describing patterns for pattern recognition, and in expressing fractals in graphic form. We will study the possibilities of handling (primary) symbol string data from genome sequences with formal grammar logic. A grammar stipulates the symbol string set and is also a tool leading to the structure of each symbol string. In the genome, it can become a tool to strictly define genome structure. By its nature, it expresses (or should express) exact features logically. It is impossible to express data statistically with the grammar itself, so we must incorporate a setup within the mechanism and fuzzy elements in order to do so. Furthermore, grammar can be a powerful means to discover high level structures efficiently in large quantities of data.

The key points in a genetic grammar description are:

- (1) Which grammar form do we choose?
- (2) How large is the object to be described?
- (3) What will be the form in which the grammar will be used?

As far as the grammar form is concerned, we can think of many forms by determining the limits of rewriting, and judging from the nature of the genome, there is no need to limit ourselves to existing grammars (such as context free grammars). As far as the size of the object to be described is concerned, we can imagine all possible options from a single sequence to a protein family to a whole gene. Concerning the form in which the grammar will be used, there is the problem of whether a human being will create the grammar (or grammar group) or whether the system (by looking at a sample) will create it automatically. Grammatical inference is the mechanism that leads to a grammar description from a (finite) set of sample symbol

sequences. Because a grammatical description stipulates a symbol sequence set, the relationship between it and the original sample set becomes a problem.

The goal of our research is to design a method for capturing genome structure automatically by paradigms of grammatical inference. For the initial step we must first determine the form of the grammar for the genome. This is expressed as the forms and limits of grammar rules. At the same time, we decide what level to make the described object (sequence set, in other words, pattern) that a grammar stipulates. Second, we create the actual grammar rules in this form for the sample described object. In addition, we also analyze the sentence structure of the sample data by these grammar rules and detect object patterns.

Details of Study

This year we first designed a framework for making grammatical inferences about the genome. For the grammatical inference format, we chose the method of replacing partial sequences of the sample symbol sequence with sequential non-terminal symbols, and for the replacements we studied the following:

- (1) Small symbol sequences that appear statistically significant,
- (2) Long symbol sequences that appear in multiple samples,
- (3) Different parts in multiple samples,
- (4) Symbol sequences having contextually identical (or similar) relationships,
- (5) Sequences within identical contexts.

To summarize the algorithm, we first are given a set of sample symbol sequences. Then we select partial sequences from the symbol sequences in the set according to one of the above types of replacements, and replace them with the same non-terminal symbol. Then we repeat the same procedure in the set of possible symbol sequences, and continue repeating until all the partial sequences are eliminated.

Second, we created a test system for structure extraction (grammatical inference) by the replacements in (1) above. In this case it is similar in form to matching identical amino acid sequences. In other words, it corresponds to the basic part of alignment. We have not performed this yet, but identical patterns will all be replaced by identical non-terminal symbols, even if they are in various locations, and they will all be recognized.

Discussion

In the future we will implement a non-terminal symbol determination mechanism by the different formats in (2) through (5), and we think these will be valuable functions. Further, we want to generalize this by joining non-terminal symbols and study a means for covering all genes that are considered members of the same family.

At present there is a large gap between the mechanism of multiple alignment in actual practice and this grammatical

inference. We wish to study how to bring these two together by learning whether alignment can be expressed by grammatical inference.

References

1. Ono, T., Takiguchi, N., Kotani, Y., and Nishimura, Y., Aminosan Hairetsu No Kozo Kaiseki, Johoshori Gakkai Heisei 4 Nen Zenki Zenkoku Takaki Koen Rombunshu [Structural Analysis of Amino Acid Sequences, Information Processing Society of Japan 1992 Biannual National Conference Proceedings], 1992.

Research on Selection of RNA Splicing Sites Using Database and Artificial Intelligence

92FE0879T Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 63-64

[Article by Hiroshi Sakamoto, Science Dept., Kyoto University, and Kenta Nakai, Basic Biological Studies Institute]

[Text]

Background and Goals

RNA splicing has immense importance in the control of gene expression of eucaryotic organisms, and in genome analysis it forms an important link that connects DNA base sequences to gene products. However, although we are gradually learning about the molecules that contribute to splicing, there are many points that remain unclear about the mechanism of splice site selection, which is most important to genome analysis. One reason this problem is difficult to solve through experimental research is that the system is too large to perform experiments directly. Moreover, the difficulty from the standpoint of sequence analysis is that we know it is probably insufficient to merely look at consensus sequences near the splice site, but we do not know what else to consider at this point. Therefore, in our research, experimental and theoretical researchers will work together, compile basic data and knowledge, pursue factors that support splice site selection, and finally compile the results we obtain into a rule-based prediction method. This year's work is the first half of a 2-year project.

Details of Study

First, we constructed a human RNA precursor database based on existing base sequence databases. We displayed the exon structure together with the splice site and consensus scores of nodes graphically as a PostScript file. In line with this we constructed a knowledge base system prototype for predicting exons from given precursors. By doing so we studied how easy it is to become familiar with the hypotheses and knowledge and with "if-then" type rules. Next, in order to study the effect of RNA secondary structure, which is a powerful candidate for a factor contributing to splice site selection, we tried to study the correlation between predicted secondary structure and the consensus score distribution. We found many new points that must be considered to incorporate existing methods into splice site prediction. Finally, we compiled data about abnormal splicing from a literature search, and we are at

present constructing an abnormal splicing database. Changes in splicing patterns accompanying point mutations and the like are thought to be a powerful clue to finding out the principles of splice site selection.

Results

First, concerning the RNA precursor database, we graphically displayed homologous sequence fragments of exon structures and splice sites for about 100 cases. Our results showed that a large number of splice site homology sequences exist within very large introns, and although there is room for improvement in the way of measuring the degree of consensus, we reconfirmed that basically the problem cannot be solved by this alone. The situation was the same even if we considered consensus sequences of nodes at the same time. However, it appeared that both ends of the fragment were relatively easy to predict. We linked this with the exon recognition hypothesis and found it very interesting. However, this hypothesis may be effective for recognizing the 5' terminus of the exon but it gives us no clue to recognition at the initial stage.

Next, we worked on the knowledge base prototype, and judging from the given precursor sequence, we tried to predict the exon fragment based on the consensus score and limitations on the length of the predicted exon. Concerning the incorporation of so-called "knowledge," in its present state some knowledge fits the "if-then" format very nicely and some does not. At present, finding the most appropriate solution in general terms from a variety of trade offs is difficult from the standpoint of expressing the knowledge, and we need to improve our use of working memory and the like.

In the case of using RNA secondary structure predictions to predict splice sites, we found many difficulties both in terms of calculating technology and biology. More specifically, it is necessary to calculate the structure of an extremely large sequence, and yet we do not know the order of secondary structure formation and the effect of protein factors. Without going into detail, we decided to use a program for predicting secondary structure written by Dr. Akiyama of the Electrotechnical Laboratory, make overlapping partial calculations, and combine the results. There are many parameters to be adjusted and in the future we will fit the effects of secondary structure into systems that have been studied and reviewed in detail.

Finally, concerning the abnormal splicing database, we have compiled over 100 references and are presently organizing them. If we recognize that the data we compiled has general importance, we plan to present this in a publication. In the same manner, we are studying a plan for compiling examples of selective splicing.

Discussion

While we were conducting our research, a splice site prediction method using a neural net was announced overseas. In that method, the prediction precision is increased using prediction results of main coding regions without worrying about the molecular mechanism of splice site selection. Of course, the precision of the prediction is far from a level sufficient for practical use. We do not

know if we will discover the principles for splice site selection in vivo in our future research. Because of many contributing factors, it will most likely be difficult to discover clear rules. It will be easy to incorporate the coding region predictions in our knowledge base. However, to continue with research for increasing prediction precision, it will truly be necessary to study a variety of experimental results and pursue rules that are closer to the actual mechanism.

References

1. Nakai, Sakamoto, Shimura, and Kanehisa, Supuraishu Bui Sentaku Ni Kansuru Ruru Hyoka Shisutemu, Dai 12 Kai Nihon Bunshi Seibutsugaku Nenkai [A System for Evaluating Rules for Splice Site Selection, 12th Annual Conference Japan Molecular Biology Society].

Study of Organizational System for Genome Analysis of Model Organism

92FE0879U Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 p 65

[Article by Hideo Shinagawa, assistant professor, Microbial Disease Institute, Osaka University]

[Text] Generally speaking, it is an important and urgent concern to develop an excellent user interface so that researchers in the field of biology who have almost no knowledge of computers can use genome analysis results and databases.

In our research we studied the ease of manipulation of GenoGraphics, which was recently developed by David G. Azwada of the Argonne National Laboratory in the United States, as an example of a system for organizing the correspondence relationships between the genes for which the base sequences have already been determined and genetic maps of the genome by using *E. coli*, whose genetic analysis and the determination of genome base sequences has progressed the farthest.

This software operates on Open Windows from Sun Microsystems, and the basic operations are performed by a mouse with almost no need to touch the keyboard.

The contents directly compare the restriction enzyme map created by Kohara et al., the genetic map created by B. J. Bachmann, the genes with determined base sequences, the base sequences between genes, restriction enzyme maps based on those base sequences and the like. This kind of software has not been developed before, and results were excellent.

However, although we were using a high performance workstation, the speed of operation lacks smoothness. In the near future we expect to be able to use workstations with higher speeds, and the operational speed problems should completely disappear in a few years.

GenoGraphics can be used by computer novices with little reference to the manual, and the presence of the UNIX is barely felt during operation. Further, we expect that a similar system will be designed with the premise that it will be used on a workstation.

Research on High-Speed Pattern Comparison Algorithm for Base Sequences

92FE0879V Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 66-67

[Article by Takeshi Shinohara, Data Engineering Dept.,
Kyushu Institute of Technology]

[Text]

Background and Goals

Comparison of character string patterns is the detection of the positions where character patterns appear within text. It has been used as the most basic and important mechanism in data search systems and database systems. DNA, the carrier of genetic data, is basically a character string composed of four types of symbols, A, T, C, and G.

The goal of our research is to address the case in which the object for pattern comparison is a DNA base sequence. We will conduct research on methods of realizing a high speed algorithm and actually construct a useful system as a tool for genome researchers.

Because of advances in the Human Genome Project, we have obtained a vast quantity of DNA base sequence data, much more than in the past. Pattern comparison is important as a basic technique for searching and utilizing this massive amount of genome data efficiently. However, almost all known pattern comparison algorithms are not designed for text with an extremely small variety of characters such as a base sequence. In our research we aim to look at speeding up pattern comparison of base sequences from a variety of viewpoints, and develop an efficient data system suitable as a tool for genome research.

Details of Study

1. High Speed Pattern Comparison Algorithm for Base Sequences

We will study a pattern comparison algorithm for base sequences from various viewpoints and explore the possibilities for increasing its speed. More specifically, we will study the applicability to base sequences of utilizing both data compression and parallel methods for greater speed.

2. Set Up SIGMA Text Database Management System

SIGMA is a general text database management system open to the general public that we developed and it is located at the Kyushu University Large Computer Center. It is an efficient data system for running high speed character string pattern comparison algorithms. We will set up this SIGMA system so that it can be used as a tool by genome researchers.

Results

1. There are very few types of characters that comprise DNA base sequence data, and almost all are A, T, C, or G. On a normal system, these are expressed unchanged by computer characters (8 bit). We studied an algorithm that compresses this data using Huffman code and scans it as it is without decoding. We learned that by doing so it is possible to compress data efficiently and perform high speed searches (References 2 and 6).

2. We converted the DNA base sequence database GenBank to SIGMA files and confirmed that we can manage and search it. If this database is opened to the public so that researchers nationwide can use it together, a variety of searches by genome researchers will be possible (Reference 3).

Future Topics

1. Transfer SIGMA System to UNIX

Because the existing SIGMA system is installed on a general large computer, there are restrictions on operating time and the like. To make this a quick responding system we will have to install it on a UNIX workstation.

2. Application to Homology Search

For base sequence data homology searches, we must have a method that uses patterns that include variables rather than a fixed pattern. The problem of automatically extracting these patterns containing variables from multiple fixed patterns that are similar, in other words the problem of extracting motifs, can be viewed as a problem of inductive inference that infers abstract rules from concrete examples. One of the ultimate goals of this research is to process a large amount of genome data more efficiently by applying a learning algorithm using inductive inference to extract general rules from these kinds of concrete examples and applying a high speed pattern comparison algorithm to those extracted patterns.

References

1. Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T., EFS No Gakushu Kanosei To Makutampakushitsu Ryoiki Yosoku E No Oyo, Johogaku Kiso [EFS Learning Capability and Its Application in Predicting Transmembrane Protein Domains, Foundations of Data Processing], 23-1, pp 1-8 (1991).
2. Fukamachi, S., and Shinohara, T., Asshuku Deta No Tame No Kosoku Mojiretsu Patan Shogo Giho, Johoshori Gakkai Zenkoku Taikai (Dai 43 Kai) Rombunshu [High Speed Character String Pattern Comparison Technique for Compressed Data, Information Processing Society of Japan (43rd) National Conference Proceedings], Vol 4, pp 83-84 (1991).
3. Shinohara, T., Arikawa, S., Kuhara, S., Miyahara, T., Inoue, H., Shinohara, A., and Uchida, C., Tekisuto Detabesu Kanri Shisutemu SIGMA To Genomu Kaiseki E No Oyo, Dai 2 Kai Kokai Wakushoppu "Hito Genomu Keikaku To Joho Kaiseki Gijutsu" Rombunshu [Text Database Management System SIGMA and Application to Genome Analysis, 2nd Public Workshop "Human Genome Project and Data Analysis Technology"], pp 29-32 (1991).
4. Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T., Negatibu Mochifu No Kikai Hakken, Dai 2 Kokai Wakushoppu "Hito Genomu Kaiseki To Johoshori Gijutsu" [Machine Discovery of Negative Motifs, 2nd Public Workshop "Human Genome Analysis and Data Processing Technology"], pp 62-65 (1991).

5. Arikawa, S., Kuhara, S., Miyano, S., Shinohara, A., and Shinohara, T., A Learning Algorithm for Elementary Formal Systems and Its Experiments on Identification of Transmembrane Domain, In Proceedings of the 25th Annual Hawaii International Conference on System Sciences, Vol 1, pp 675-684, January 1992.

6. Fukata, S., Shinohara, T., and Takeda, M., Kahen Cho Fugo Asshuku Deta No Tame No Mojiretsu Patan Shogo—Genomu Joho No Kosoku Kensaku Giho, 1992 Nen Johogaku Shimpojium [Character String Pattern Comparison for Variable Long Code Compressed Data—Method for High Speed Search of Genome Data, 1992 Information Science Symposium], pp 95-103 (1992).

Development of UNIX Shell for Genome Data Analysis

92FE0879W Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 68-69

[Article by Akira Sueyama, Engineering Dept., Technological University of Nagaoka]

[Text]

Background and Goals

Genome data analysis can be roughly divided into routine analysis, in which the analysis is widely used and clear cut, and non-routine analysis, in which the contents of analysis are quite varied and change dynamically. To proceed efficiently with research on genome data analysis, a system suited to non-routine analysis processing is effective. In routine analysis systems as well, a system and analysis environment based on a non-routine analysis system is desirable for responding in a timely manner to analysis of data that changes rapidly.

The UNIX system has already received high marks as an environment for proceeding efficiently with the development of software that is a model for non-routine operations. On the UNIX, the separate tools and modules provide as much universality, independence, and potential for reuse as possible, and it adopts the concept that complex processing is realized by combining these functionally. This kind of concept is suitable for performing non-routine analysis processing of genome data that is quite varied and dynamically changing. Based on the proven software techniques and concepts, in our research we designed and developed a Genomic Shell (GSH) that is an environment suitable for proceeding efficiently with non-routine genome data analysis on the UNIX.

Results

To divide up complex genome data analysis processing into multiple processes and perform it, there must be an exchange of data by communication between the processes. With the GSH we developed in our research, we used a client server module to perform communication between processes by SOCKET and SYSTEM CALL.

The server of the GSH was developed as a data management function located in the memory and a function that analyzes large amounts of genome data at a high speed.

The object, which is a collection of data within the server, has a hierarchical structure to perform operations efficiently. These operations and system management are performed in the same manner as in the UNIX file system. As forms of data objects, it supports both variable length data objects (suitable for the storage of data in which the size changes with the circumstances, such as base sequence data) and fixed length data objects (suitable for storing data in which the size is constant, such as molecular structure coordinate data) and data that must be accessed at high speeds. As data analysis functions are concerned, on the server it particularly supports analysis in which high speed processing is required and analysis in which the implementation becomes quite slow in the client process because of the necessary exchange of large amounts of data between processes.

For the standard client process, we designed and developed a basic process library for displaying the GSH server operation, analysis of basic genome data, and the data object within the server. To maintain continuity of the data analysis processing operation on GSH, analysis processing of the basic process library all runs in the form of operations on the data object. Concerning the details of the data object, for example, we can see an image by creating that image on a window through appropriate mapping conversion, and this mapping conversion is also created as a data object. If a data object whose details you want to display is processed according to a mapping conversion data object using the GSH universal formatter FMTG, it creates a display matching those details.

Standard genome data analysis with GSH is performed by combining functions such as the UNIX pipe, redirection shell script and the like with these basic processes, the abundant tool library supported by the UNIX, free software, and a process library extended by the user for various analysis processes. The user can easily extend those functions on the GSH by saving an analysis processing procedure that he has created as a user extended process library.

The standard method for utilizing the GSH environment is on the process level, but we supported GLIB programming functions and the GSH server extension function GSET for higher level utilization. GLIB is a C language-related library for using the GSH server functions. By using GLIB relations in the C language program created by the user, it is possible to use the server function at a higher speed and with very great precision. To extend the GSH functions the server function itself must be extended. GSET is the tool for doing so. It is possible to add user-created analysis program functions to the server by using GSET.

GSH is a system that emphasizes the extendability and flexibility required for non-routine data analysis processing rather than manipulative capability. However, manipulative capability is very important for GSH to be widely utilized by general users who are not familiar to computers. Therefore, we tried to operate GSH in a GUI environment using VIOLA, which is an interpreter-type

object-oriented language. We found that we could use GSH in the GUI environment by utilizing VIOLA as one of the client processes.

Discussion

This year we designed and developed the most basic part of the GSH system. Therefore, the server functions and basic process library supported this year are very limited, and not adequate for performing actual genome data analysis. The GSH system is one with self-extension capability. In the future we want to realize the full capabilities of the basic functions of the server and the basic process library by using this function, prepare a manual, and perfect an environment in which actual genome data analysis can be performed.

Research on High-Level Processing of Protein Amino Acid Sequence Data Based on Pattern Recognition Methods

92FE0879X Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 70-72

[Article by Yoshimasa Takahashi, Motokazu Kamimura, and Shinichi Sasaki, Engineering Dept., Toyohashi University of Technology]

[Text]

1. Background and Goals

It has been demonstrated in many experiments in the past that the amino acid sequence data of proteins contains both the transcription information from the gene and a variety of information that stipulates the high level structure and function of the protein itself, which are its ultimate form of expression. As analytical techniques have improved in recent years, primary sequence data has increasingly accumulated, and the quantity has already become massive. However, data concerning the high level structure and functional expressions reflected in this primary sequence data have not accumulated in sufficient amounts, partly because of the difficulty of the associated analytical techniques. Therefore, we need to establish a powerful methodology for dealing with existing data to acquire the techniques for analyzing the various types of information hidden in amino acid sequences and the knowledge linking it to high level data.

Our research aims to establish a method for high level analysis of sequence data based on the use of pattern recognition methods by analyzing the correlations between this amino acid sequence data in proteins and their high level structure and function.

2. Details of 1991 Research

Until now the classification based on the so-called secondary structures of α -helix, β -sheet, hairpin turn and the

like has been widely used as the classification of characteristic, localized three dimensional structures of the protein backbone. However, these secondary structure classifications describe rather large, characteristic, localized structures consisting of up to 20 continuous residues. On the other hand, the establishment of a new methodology for analyzing these data and a more detailed analysis of protein structural features will be required to clarify the formative principles of protein high level structure when trying to predict high level structure from the amino acid sequence. Therefore, in our research we attempted protein amino acid residue ϕ - ψ conformational pattern (called ϕ - ψ pattern below) clustering by the potential function method with the goal of classifying the local three dimensional structures of specific residues from a viewpoint different from that of traditional secondary structure classification.

3. Method

Data Set: We calculated the dihedral angles ϕ and ψ of the backbone from the three dimensional data of known three dimensional structures of 97 proteins (21,941 residues) listed in the PDB (Protein Data Bank). We can think of this dihedral angle data as a single pattern point on a ϕ - ψ two dimensional plot. In other words, pattern X is expressed as $X = (\phi, \psi)$.

Cluster Analysis: First we expressed the density distribution of each pattern in the ϕ - ψ pattern set, which is the object of clustering, with the Gauss distribution type potential function proposed by Coomans and Massart [Reference 1]. Then we made the pattern with the largest average potential function value the initial clustering pattern, and increased the class by taking each pattern, one by one, that approached this initial pattern and classifying it either as belonging to the same class as the initial pattern or to a different class. Here the clustering was according to the nearest neighbor method. When the growth of the cluster containing the initial pattern was completed, we made a new ϕ - ψ pattern set excluding all patterns included in that cluster. We repeated the same operation either until all classes to which patterns belonged had been determined or until the predetermined ending condition (largest cluster number) had been satisfied.

4. Results and Discussion

We created a ϕ - ψ conformation pattern distribution map for each residue of the 20 amino acids using the potential function method. As one example, the distribution map for the ϕ - ψ conformation pattern of Gly residues is shown in Figure 1. Based on these data we performed mode searches of conformation patterns in ϕ - ψ space and cluster analysis. In every case we detected two or more major clusters. More specifically, we discovered five main clusters from the ϕ - ψ patterns with Gly (1,991) as the center residue, and we learned that about 70% are included in one of these five clusters. Table 1 shows the dihedral angle (ϕ, ψ) pattern at the center of each cluster based on the mode search of these Gly residues and the number of bases the pattern consists of.

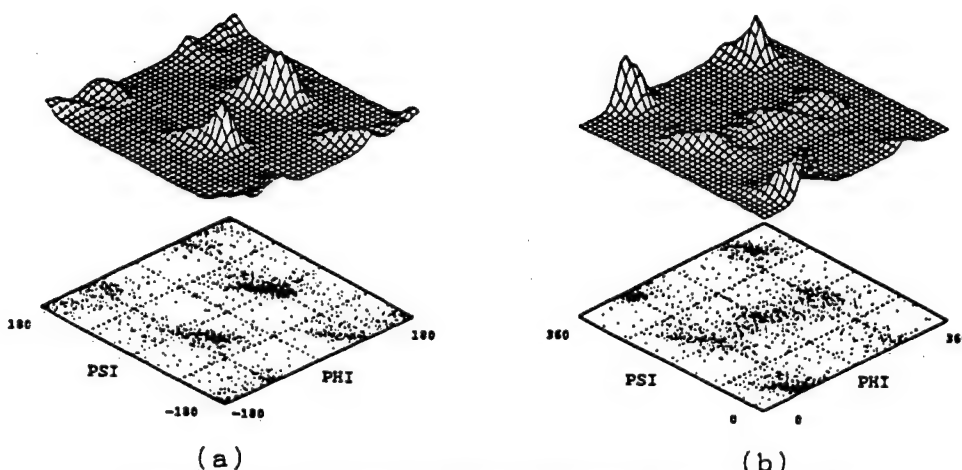


Figure 1. ϕ - ψ Conformation Pattern Distribution Map of Gly Residues

(a) Map drawn round $\phi = 0^\circ$, $\psi = 0^\circ$; (b) Map drawn round $\phi = 180^\circ$, $\psi = 180^\circ$

Table 1. ϕ - ψ Conformation Pattern Clustering of Gly Residues

| Cluster No. | ϕ | ψ | N |
|-------------|--------|---------|------|
| 1 | 85.07 | 5.58 | 511 |
| 2 | -61.98 | -42.67 | 317 |
| 3 | 84.19 | -160.72 | 159 |
| 4 | 176.80 | -170.03 | 220 |
| 5 | -79.58 | 168.65 | 155 |
| Others | — | — | 629 |
| Total | | | 1991 |

Further, when we performed the same kind of analysis on other amino acid residues, we detected two or three main clusters, and an average of about 80% of the residues were included in one of the clusters.

This research proposed a method for mode searching and clustering using potential functions for the data analysis of protein backbone local structural information, and we clarified the main clusters for each amino acid by applying this method to the analysis of ϕ - ψ conformation patterns of each of the 20 amino acids. We believe that this enables a classification of protein local structural features that differs from traditional secondary structure classification, and at the same time, it offers new data on discovering the relationship between the local three dimensional structure of the protein backbone and its amino acid sequence.

References

1. Coomans, D., and Massart, D. L., Anal. Chim. Acta, 133, 225-239 (1981).

Publications

1. Kamimura, M., Takahashi, Y., and Sasaki, S., ϕ - ψ Conformational Pattern Clustering of Protein Amino Acid Residues Using the Potential Function Method, J. Protein Chemistry, submitted.

Application of Case Based Reasoning to Intelligent Processing of Genome Knowledge Database

92FE0879Y Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 73-74

[Article by Takano Terano, management systems science specialist, Tsukuba University]

[Text]

Background and Goals

In the intelligent processing of genome data, it is important to establish (1) techniques for storing and searching vast quantities of data, (2) techniques for handling the requisite knowledge about genome data analysis, and (3) techniques for effectively using past analysis experience. We conducted research with the goal of integrating (1) and (2), mainly from the standpoint of (3). We used case based reasoning (below CBR) as our basic technique. CBR is a new inferential method in which past problem solving examples are stored in the computer, and when a new problem is presented, it uses those actual cases (successes and failures) to solve the problem by making analogies to those past cases. CBR is based on analogy as a technique of artificial intelligence, and lies between database technology and deductive inference technology. It has a strong connection to inductive and deductive knowledge acquisition technology.

The goal of our research is to propose a technique for integrating a database that can store and use massive quantities of genome data relatively easily with a knowledge base that provides knowledge that is deeply related to genome data analysis in a system based mainly on CBR. The basis for this method will be the realization of flexible functions that enable efficient intelligent searches of the database to meet the routine requirements of users, and at

the same time, utilize vast knowledge and past problem solving experiences to meet the requirement of solving complex, difficult problems.

Details of Study

This year we conducted the following three studies on the applicability of artificial intelligence technology based on CBR from a general point of view. We mainly used the LISP machine board (MacIvory) that we purchased this year.

(1) Study of Methodology for Development of Knowledge System

We conducted research on methods of developing a large scale knowledge system having the various functions of a database, a math processor, and a knowledge base, and we studied the practicality of such a system structure [Terano 1991].

(2) Study of a Knowledge System Equipped With a Learning Function

We performed a study on methods of efficient problem solving by applying a deductive learning process called Explanation-Based Learning (below EBL) for a knowledge system equipped with a vast knowledge of the subject area [Ogata 1991].

(3) Study of Case Based Reasoning Trade-Offs

When CBR functions are incorporated into an actual system, it cannot cover the whole solution space with actual cases, so it must coexist with a traditional problem solving system. As a result, in a CBR system there are cases in which the problem solving capability actually drops because the load falls on processing the cases. Therefore, we considered trade-offs of the CBR system, and performed a study of the conditions in which its effectiveness can be displayed.

Results

The main results obtained in the above three studies are summarized below:

(1) For the development of a large scale knowledge system, it is important at the outset to identify the applicable level and scope of artificial intelligence technology. It is also important to capture and approach the knowledge acquisition activity from the standpoint of the life cycle of system development.

(2) EBL will be useful for improving processing efficiency provided sufficient knowledge can be incorporated into it beforehand. It will be necessary to incorporate this as part of the CBR system.

(3) There are the following four types of CBR trade-offs. 1) Difficulty of case recoverability/difficulty of direct problem solving, 2) sufficient problem solving function capability/sufficient case base capability, 3) sufficient number of cases/case search efficiency, 4) index knowledge of cases/knowledge acquisition for that purpose. Based on appropriate assumptions, it will be possible to understand

the performance of the CBR system as a sum total of these trade-offs, and actually evaluate the performance of the system.

Discussion

This is our first year of research therefore, we conducted research from the general position of artificial intelligence technology. Genome data processing differs from the topic of our research because of the specialized knowledge required for processing genome data, but it has much in common with artificial intelligence technology. In building of a large scale system such as in our research it is essential to separate the part that requires a large amount of specialized knowledge for problem solving from the part that enables efficient manipulation of large quantities of data, and then integrate them holistically. CBR is effective as a basic technique for doing so.

References

1. Nabeta, S., and Terano, T., Jireibesu Suiron Kiko No Koritsu Ni Kansuru Bunki To Hyoka, 1991 Nendo Jinkochino Gakkai (Dai 5 Kai) Zenkoku Taikai Rombunshu [Analysis and Evaluation of Efficiency of Case Based Reasoning Mechanism, 1991 Proceedings of (5th) National Conference of Artificial Intelligence Society], 3-2, pp 223-226, 1991.
2. Ogata, T., and Terano, T., Explanation-Based Narrative Generation Using Semiotic Theory. Proc. Natural Language Processing Pacific Rim Symposium (NLPRS '91), pp 321-328, 1991.
3. Terano, T., Chishiki Shisutemu Kochiku Hohoron Ni Kansuru Kenkyu—Suiryokuko Kozobutsu Hyoka Shindan Shisutemu no Kaihatsu Ni Sokushite—Tsukuba Daigaku Keiei Shisutemu Kagaku Senko Resachi Repoto [Research on Intelligent System Construction Methodology—Development of Diagnostic Evaluation System for Hydraulic Steel Structures, Tsukuba University Management System Science Research Report], No 1991-01, 290 pp (submitted to Tokyo Institute of Technology as Ph.D. Dissertation).

Description Method for Protein Three-Dimensional Structure by Logical Type Language
92FE0879Z Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 75-76

[Article by Toshiyuki Noguchi, Science Dept., Nagoya University]

[Text]

Background and Goals

Knowing the three-dimensional structure of proteins is an important problem in understanding the meaning of genetic data. Analysis of genetic data must include not only the relationship between a base sequence and the primary structure amino acid sequence, but also the relationship between primary structure and three-dimensional structure, and the relationships between different three-dimensional structures. To analyze a relationship there must be a method to express both sides. Base sequences

and amino acid sequences can be described without manipulation in one dimension, but a method to describe the complex, three-dimensional structures of proteins has yet to be established. Generally speaking, a structural description consists of the description of the various components and their relationships. This means that the logic language PROLOG, which uses a Horn clause format of first order predicate theory to describe relationships, is suitable as a computer language to describe the three-dimensional structures of proteins. In 1985, A. J. Morffew and S. P. J. Todd developed an inquiry system for protein structure using the logic language PROLOG in place of a relational database. This was only an experiment, but it demonstrated the potential of PROLOG as a structural description language. Moreover, T. Takagi et al. developed an inferential database system that can query with logic programming, and applied it to a system for searching protein three-dimensional structures. The goal of our research is to clarify the various problems in establishing a method for describing protein three-dimensional structures required for the computer analysis of genetic data, and to build its basic foundation.

Outline of Basic Specifications of Description

[1] There are three types of items to be described: data, components, and relationships.

Data: Data on protein structure. There are the following two types.

- 1) Individual data on each protein. Examples: atomic coordinates, amino acid sequences, S-S bonds, etc.
- 2) Data common to all proteins. Examples: attributes of atoms and amino acids (charge, hydrophobicity, size, etc.).

Components: Structures are expressed as relations between components. There are the following four types of components.

- 1) Basic components: atoms, residues, polypeptide chains, proteins, amino acids, atoms comprising amino acids, others.
- 2) Partial chains: parts of polypeptide chains.
- 3) Structure: expressed as relationships between components.
- 4) Sets of components.

Relationships: There are the following three types.

- 1) Relationships between components and their attributes. Examples: relationship between atoms and their coordinates.
- 2) Relationships expressed as structures. Examples: hydrogen bonds are expressed as two atoms and a connection between them, and this is a structure.
- 3) Relationships between the attribute values and attribute value relationships and their coefficients. Example: relationship between size and charge of atoms.

[2] Methods of Expression

Data: The relationships between basic components and their attributes are expressed as facts.

Components:

- 1) Basic components are expressed as (PROLOG) atoms.
- 2) Partial chains are expressed as compound items.
- 3) Structures are expressed as compound items.
- 4) Component sets are expressed as lists.

Relationships: Expressed as predicates.

Definitions of structures: Expressed as rules.

Specifications for Basic Predicates

Among the most basic relationships, the relationships between atoms, residues, polypeptide chains, and proteins and their attributes are expressed as the following predicates.

```
atom(Atom, Atom_name, X, Y, Z, Res)
res(Res, Amino_acid, Res_no, Chain)
chain(Protein, Chain_id, Sequence_list, Chain)
protein(Protein, Protein_name, Class, Source)
```

The atomic coordinates from the Protein Data Bank, etc., are expressed as facts in the above predicate format. In addition, shared data such as amino acid attributes are also expressed as facts and added to the database. We also will add basic items from the predicates that express relations between attribute values.

Discussion

Time is required to search out desired facts that are written in basic predicates from the database. First we must solve this problem. The actual content of the analysis of three-dimensional structures is a relative classification of protein partial structures. The main operation is searching out structures defined by rules. In the future we will define rules with basic predicates and describe structures such as hydrogen bonds and secondary structure. We will then define high level structures by the rules that use predicates expressing these structures. While performing analysis of three-dimensional structures by using this descriptive method we will study what kinds of predicates and structures show useful relationships and elements for description.

References

1. Yoshikawa, K., Noguti, T., Tujimura, M., Koga, H., Yasukochi, T., Horiuchi, T., and Go, M., Hydrogen bond network of cytochrome P-450cam: a network connecting the heme group with helix K. B.B.A. (in press).

Construction of Major Histocompatibility Complex DNA Database and Development of Visual Software

93FE0334A Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 77-78

[Article by Hiroshi Hori and Masaaki Matsuura, Institute of Radiological Medicine for Atomic Bomb Exposure, Hiroshima University, and Kazuo Yoshida, Science Dept., Hiroshima University]

[Text]

Background and Goals

The major histocompatibility complex (MHC or HLA) has been emphasized in the past as a transplantation immunity antigen during organ transplants, and it is an important protein group in day-to-day clinical operations because HLA typing is done to determine transplant donors. On the other hand, MHC molecules exhibit "abnormal" genetic polymorphism as constituents of the organism, as demonstrated by the existence of the large number of HLA forms. This polymorphism derives not only from polymorphism near each gene locus but from multiple gene loci as well. It is estimated that there are several hundred in the human group alone. However, such data is still incomplete.

This research plan proposes to extract data on the complex MHC gene family from gigantic databases such as GenBank, and create a new, more efficient, easier-to-use, basic supplemental database and a search and analysis tool. In addition we will create a computer tool to express and process those analytical results, and eventually, create an integrated WINDOW environment for BIOSOFT. This software and database will not only provide basic data for clinical HLA typing, but will also be useful in research in molecular biology and applied biology including biotechnology, which is based on molecular biology, and in education. For these reasons, we believe it will have an extremely large impact.

Details of Study

DNA databases such as EMBL and GenBank can be used on large computers and on microcomputers as essential basic fact databases in biology and biotechnology. However, a genetic database and software to search it have not yet been developed. Recently, we developed a prototype software called GENEMAP for use on microcomputers for building and searching genetic databases. With GENEMAP it is possible to search for genes from a database on a color graphic chromosome map, display them, and arbitrarily call up and display DNA base sequences, amino acid sequences from a protein database, and literature references. We plan to transfer this software to a UNIX system workstation that has more powerful graphic functions and provide it to researchers.

Results

We are proceeding with the creation of a human MHC gene family database consisting of an equivalent number of GenBank data and original research articles. Partial results of analysis using this database were presented at the Human Genome Project Symposium, International Association of Human Biologists Conference on Isolation and Migration (in Fukui, Japan) sponsored by the Japan Society of Human Genetics. Moreover, in 1992 at the group conference for this area of research, we introduced an expanded database and presented results of analysis using it. We also developed a prototype search and construction software for genetic databases called GENEMAP for microcomputers, and by using this it is possible to search and display data from genetic databases on a color graphic chromosome map. Moreover,

we can freely call up and display the DNA base sequences, amino acid sequences from protein databases, and literature references.

Discussion

It is necessary to transfer this human gene family database and its software to a UNIX workstation with more powerful graphic functions, expand those functions, improve it, provide it to researchers, and expand its human interface.

References

1. Hori, H., and Satow, Y., Dead-end evolution of the Cnidaria as deduced from 5S ribosomal RNA sequences. *Hydrobiologia* 217: 505-508 (1991).
2. Hori, H., and Satow, Y., Archaeobacteria vs. Metabacteria: Phylogenetic tree of organisms indicated by comparison of 5S ribosomal RNA sequences. In: "Evolution of Life" (Eds.: Osawa, S., Honjo, T.), pp 325-336, Springer, Tokyo/Berlin (1991).
3. Hori, H., Molecular phylogeny of organisms and cellular systems. In: "Biological Diversity and Global Change" (Eds.: Solbrig, O. T., Van Oordt, W., Van Emden, T.), in press (1992).
4. Hoshino, S., Hori, H., et al., PCR-SSCP analysis of HLA class II DRB1 and DQB1 genes: A simple rapid method for histocompatibility test. Transplantation (submitted).
5. Okamoto, K., Suzuki, K., and Yoshida, K., Physical mapping and RFLP analysis of mtDNAs from the ascosporogenous yeasts: *Saccharomyces exiguus*, *S. kluyveri* and *Hansenula wingei*. *Jpn. J. Genet.*, 66, 709-718 (1991).
6. Nishikawa, M., Suzuki, K., and Yoshida, K., DNA integration into recipient yeast chromosomes by transkingdom conjugation between *Escherichia coli* and *Saccharomyces cerevisiae*. *Current Genet.*, 21, 101-108 (1992).
7. Hori, H., Heni No Bunshi Shinka: 5S rRNA to Hito Shuyo Soshiki Tekigo Kogen Idenshigun, Gekkan Kaiyo [Molecular Evolution of Mutations: 5S rRNA and Human Histocompatibility Complex Gene Family, Monthly Ocean], 23, 103-113 (1991).

Research on Acquisition of Knowledge From Mass Quantities of Genome Data by Parallel Learning Algorithm

93FE0334B Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 79-80

[Article by Satoru Miyano and Setsuo Arikawa, Science Dept., Kyushu University]

[Text]

Background and Goals

Human DNA is a huge information molecule consisting of a total of 3×10^9 base pairs, and all human genetic information is described therein. There is a strong demand both from a biological and a sociological standpoint for this information to be decoded. One of the ultimate goals of genome analysis is to extract high level information such

as the blueprint of life and information concerning the regulation of expression from the vast quantities of genome data that accumulates. Therefore, the purpose of this research is to construct a process for acquiring the knowledge hidden within gene and protein data as symbol strings through the use of various learning algorithms.

Details of Study

To construct this knowledge acquisition form we studied the following three items.

1. Capture the process of knowledge acquisition from genome data as a language learning framework that infers the grammar controlling that process.
2. Perform a theoretical analysis of the quantity of calculations in the process of analogous learning and parallel processing.
3. Perform theoretical and experimental analyses of the process of knowledge acquisition from genome data by new learning paradigms such as stochastic approximation learning.

Results

We obtained the following results from the above studies.

1. We formulated knowledge acquisition from genome data as an EFS (Elementary Formal System) learning process, which is a type of logic program. We identified polynomial-time learnable EFS partial classes on the basis of stochastic learning via theoretical research, and discovered a polynomial-time learning algorithm. However, even though we say it is polynomial-time learning, because this time is not realistic, we added several theoretical and rational restrictions, introduced an approximation algorithm, and tested it as a system that can actually be used. Then we performed an experiment to identify membrane-spanning domains of proteins using the PIR database and obtained satisfactory results. Moreover, this learning algorithm showed that can be made parallel in principle.
2. We demonstrated that an NP-complete problem situation appears in the Arikawa-Haraguchi analogy mechanism through research on analogy.
3. We considered knowledge acquisition from genome and protein data given as symbol strings and created the concept of a decision tree over a regular pattern. Then we constructed a classification and learning form according to the decision trees over regular patterns. This is based on the principle that a given sample is rationally explained by a small hypothesis in a form that learns decision trees over regular patterns of interest that are classified as positive examples and negative examples from a randomly chosen sample. We tested an experimental system based on this learning form and performed calculation tests using the

same material as in 1. above. This machine learning system extracted motifs from portions other than membrane-spanning domains in membrane-spanning domain identity problems, and we discovered a hypothesis for describing with a precision of 95% or better membrane-spanning domains and other portions. We tested the membrane-spanning domain prediction program, which is based on the knowledge we discovered, and obtained very good results.

Discussion

The above process was constructed entirely while we made deep, theoretical observations on computer learning and obtained results. The system we tested gave very good results in tests concerning identification of membrane-spanning domains. In this manner we were able to confirm to a certain degree of satisfaction the usefulness of these knowledge acquisition forms from theoretical and experimental observations of them, and we firmly believe that these forms are extremely useful for advancing this research. The task that remains is to confirm experimentally what kind of knowledge acquisition these forms are useful for, and we eagerly await the results.

On the other hand, many problems that must still be solved appeared in the theoretical and experimental results on the forms we developed and studied. In the future we must consider using parallel processing, making practical studies on forms according to analogy, and doing theoretical research on learning forms by decision trees over regular patterns.

We also learned during this year's research that a fully capable computer environment such as a workstation, disk server, network and the like will be necessary to address this research head on and expand it.

References

1. S. Miyano, A. Shinohara, and T. Shinohara, Which classes of elementary formal systems are polynomial-time learnable? Proc. 2nd Algorithmic Learning Theory, 139-150 (1991).
2. S. Furuya and S. Miyano, Analogy is NP-hard. Proc. 2nd Algorithmic Learning Theory, 207-212 (1991).
3. S. Arikawa, S. Kuhara, S. Miyano, A. Shinohara, and T. Shinohara, A learning algorithm for elementary formal systems and its experiments on identification of transmembrane domains. Proc. 25th Hawaii International Conference on System Sciences, Vol I, 675-684 (1992).
4. S. Arikawa, S. Kuhara, S. Miyano, Y. Mukouchi, A. Shinohara, and T. Shinohara, A machine discovery from amino acid sequences by decision trees over regular patterns. RIFIS-TR-CS-44, Research Institute of Fundamental Information Science, Kyushu University (1991).

Development of Artificial Intelligence System for Genetic Data Analysis Based on Molecular Evolution

93FE0334C Tokyo MOE KEY AREA RESEARCH
in Japanese Mar 92 pp 81-82

[Article by Tamio Yasukawa and Ryoichi Kataoka, Engineering Dept., Tokyo University of Agriculture and Technology]

[Text]

Background and Goals

To clarify what special functions of proteins coded by genes and their associated sugar chains, ions and the like, it is important to understand from the viewpoint of molecular evolution the ways these structures and functions have changed and developed. This research aims to develop an integrated system for analysis with emphasis on how protein three-dimensional structures, which have a decided effect on protein function, have developed by molecular evolution. This is impossible simply by the calculation of numerical values, so our research centers around a database on the evolutionary development of correlations between partial structures and functions, and an artificial intelligence system that can make flexible inferences to analyze this data.

Details of Study

As the first step in reaching the above goal, we began development of a system to predict what kind of high level structure a polypeptide chain with a specific primary structure will have. Many attempts have been made to solve the problem of predicting this high level structure with little success. This indicates that this problem is impossible to solve by using only one method. Therefore, in our research we used different models on four different levels: (1) the relative state of denaturation, (2) the phase of secondary structure formation, (3) molten/globular phase, (4) determination of the optimal structure at the atomic level.

In a relatively hot water bath, a polypeptide chain moves in large segments. In most cases the conformation is governed by entropy factors. Based on this assumption, we devised a pearl necklace model by joining ring components that participate in hydrophobic interactions to the repulsion potential around the alpha-carbon atom in each residue of the peptide chain via hypothetical bonds 3.8 Angstroms in length. The conformation of this model chain was determined from the entropy resiliency while considering the excluded volume effect of the components, in addition to the above two types of potentials. We evaluated the repulsion diameter for each residue that was needed for the simulation from PDB data and used experimental values for the hydrophobicity parameter. We introduced the segment movement as a crankshaft type movement based on random values. Using this model chain, we repeated a simulation that recreated the process of thermal annealing in which the temperature was lowered from 100°C to 27°C with an effective conversion value of 5×10^4 times.

Results

We performed a folding simulation from the denatured state for BPTI, which is comprised of 58 residues, using the above pearl necklace model. We discovered that as the folding progressed, the hydrophobic residues concentrated in the interior and that both the distance between the ends of the chain and the diameter of inertia, although fluctuating greatly, gradually approached the value of the crystalline structure. In the crystal, three S-S bonds are formed, and this is comparable to experimental data based on the possibility of S-S bond formation and variations in the distance between the six Cys residues during the folding process. At this point we are not at the stage of reaching a final conclusion, but we obtained several interesting results, and we expect that a quantitative discussion about the experimental results of the folding process of BPTI will be possible through a revision of the parameter values and increased precision of the model. A distance map is useful for comparing the relative positions of residues in three-dimensional space. In the simulation process distances on the map fluctuate greatly, but we discovered that the map of distances between alpha-carbon atoms approaches that of the crystalline structure when the Cys-Cys pairs are near their crystalline values. The fact that a three-dimensional structure close to that of a crystal appears in a simplified model such as the pearl necklace model has major implications in discussing the factors for high level structure formation, and a very careful study of this will be needed in the future.

Discussion

We created a neural network system for predicting secondary structure by back propagation but could not obtain sufficient results. We can say that one reason may be the following. We know the crystalline structures of only about 400 proteins. However, judging from the fact that the data of more than 20^{10} partial structures would be needed if we say, for example, that each residue is affected by the 10 residues ahead of it and behind it, we do not have a sufficient quantity of data for statistical estimation of secondary structure formation. Moreover, in past prediction methods only the effect of residues along the sequence was taken into account, but based on the fact that the effects of other residues that are spatially nearby is also important especially in the stabilization of β -sheet structures, we believe a sweeping modification of secondary structure prediction is needed. Therefore, we recommend the development of a system based on the Hopfield model. In this model the conformation energy of the polypeptide chain is given as a balance of interactions between the residues, and because the interaction between residue i and $i + k$ is governed by the intermediate bonding forms of $i + 1, \dots, i + k - 1$, if we can create a database of the energies of interactions between residues in each major partial structure and combine them by neural network, we will be able to solve local conformation problems of peptide chains as generalized optimization problems. At present, we are building a database of the energies of interactions between residues for this purpose.

References

1. T. Karasawa, K. Tabuchi, J. Izumisawa, and T. Yasukawa, "Development of simulation models or protein folding in thermal annealing process I." Computer Applications in the Biosciences, submitted for publication.